Approches pénalisées et autres développements statistiques pour l'épidémiologie

Vivian Viallon

Université Claude Bernard, Lyon 1

Habilitation à Diriger des Recherches

présentée par

Vivian Viallon

Approches pénalisées et autres développements statistiques pour l'épidémiologie

Soutenue le 24/05/2016
devant le jury composé de

A. Chambaz	Univ. Nanterre	Rapporteur
D. Commenges	INSERM, Bordeaux	Examinateur
A-L. Fougères	Univ. Lyon 1	Présidente
V. Rivoirard	Univ. Dauphine	Rapporteur
S. Robin	INRA, Paris	Examinateur
R. Thiebaut	Univ. Bordeaux, CHU Bordeaux	Rapporteur

Table des matières

Avant-propos

Ce document de synthèse résume les travaux que j'ai effectués ou initiés depuis septembre 2009, qui correspond au début de mon séjour post-doctoral à l'Université de Berkeley. C'est également à partir de cette date que j'ai commencé à travailler sur les approches pénalisées, qui constituent aujourd'hui une part importante de mes activités de recherche. J'en profite pour remercier Laurent El Ghaoui et Bin Yu de m'avoir accueilli pour une année au sein du projet StatNews. Je tiens également à remercier ici Joël Coste de m'avoir préalablement accueilli au sein du service de biostatistique de l'hôpital Cochin pendant trois ans à l'issue de mon doctorat et d'avoir initié le projet autour des modèles graphiques binaires, où les approches pénalisées ont fait leur première apparition pour moi.

A travers les postes que j'ai pu occuper, j'ai souvent été au contact direct de cliniciens ou d'épidémiologistes : pendant ma thèse que j'ai effectuée en grande partie au sein de l'équipe INSERM E3N dirigée par Françoise-Clavel-Chapelon, puis lorsque j'étais Assistant Hospitalo-Universitaire au service de biostatistique de l'hôpital Cochin et de l'université Paris Descartes, et enfin depuis mon arrivée à l'UMRESTTE (Unité Mixte de Recherche Épidémiologique et de Surveillance Transport Travail Environnement). Cette proximité m'a conduit à réaliser différents travaux purement applicatifs, et m'a ainsi permis de me familiariser avec des problématiques plus ou moins spécifiques de l'épidémiologie. Ces travaux applicatifs ont aussi été une source d'inspiration et ont finalement guidé la plupart de mes travaux méthodologiques voire théoriques de ces dernières années.

Le chapitre introductif de ce document présentera succinctement certaines de ces problématiques, et les questions d'ordre méthodologique qu'elles ont soulevées. Nombre d'entre elles s'interprètent comme un problème de sélection de variables. Celui-ci est des plus classiques en statistique, et des approches dérivées de critères pénalisés sont connues pour pouvoir le résoudre sous certaines hypothèses. Sous des modèles paramétriques, ces approches encouragent des structures particulières dans le vecteur des paramètres telles que la parcimonie ou l'égalité de certaines composantes, etc. Dans la première partie de ce manuscrit, je présente des résultats généraux sur des approches pénalisées par la norme L_1 des paramètres ou des dérivées de cette norme. La seconde partie est quant à elle consacrée à mes travaux sur l'utilisation de ces normes dans un contexte particulier, que je qualifie de données stratifiées. Dans ce cadre, une des questions principales est de déterminer si le niveau d'association entre deux variables est identique chez tous les individus d'une population ou si au contraire il varie à travers des sous-groupes prédéfinis de cette population

(ou strates).

Dans la dernière partie, je présente des travaux sans doute plus spécifiques encore à l'épidémiologie et à la recherche clinique. Par souci de concision, j'ai décidé de me concentrer sur mes travaux récents relatifs à l'inférence causale, et de ne pas présenter ceux concernant l'évaluation des modèles pronostiques et des tests diagnostiques.

Je vais conclure ce très bref résumé comme je l'ai commencé, par des remerciements. Je tiens tout d'abord à remercier Antoine Chambaz, Vincent Rivoirard et Rodolphe Thiébaut pour avoir accepté d'être les rapporteurs de mon HDR, et aussi Daniel Commenges, Anne-Laure Fougères et Stéphane Robin pour avoir accepté de participer au jury de soutenance. Je remercie également Bernard Laumon, Jean-Louis Martin et l'ensemble des membres de l'UMRESTTE ainsi que les membres de l'Institut Camille Jordan (en particulier, et une nouvelle fois Anne-Laure) pour leur accueil : travailler dans un tel environnement est clairement précieux. Mon intégration dans le paysage lyonnais doit beaucoup aussi à Franck Picard, qui est de plus source de nombreux conseils avisés. J'en profite pour remercier l'ensemble de l'équipe Statistique en Grande Dimension pour la Génomique du Laboratoire de Biométrie et Biologie Evolutive, qui m'accueille régulièrement dans son groupe de travail. Je remercie de même René Ecochard, Laurent Jacob, Delphine Maucort-Boulch, Nelly Pustelnik, Muriel Rabilloud, Pascal Roy et Fabien Subtil avec qui j'ai la chance d'enseigner au sein du Master de Santé Publique ou du M2 Maths en Action. Un grand merci aussi à Pietro Ferrari, Sophie Lambert-Lacroix, Aurélien Latouche, Grégoire Rey et Adeline Samson pour des collaborations enrichissantes, ainsi qu'à Philippe Rigollet qui sait toujours trouver du temps, notamment pour répondre à mes questions techniques de dernière minute. Et bien sûr merci aux étudiants que j'ai encadrés en thèse ou en stage : Edouard, Marine, Nada, mais aussi Alexei, Cécile, Yacine , etc. J'espère avoir réussi à vous transmettre quelques compétences ; dans tous les cas, votre motivation a été un moteur pour moi.

Enfin, et évidemment, mes dernières pensées vont à Virginie et Lucile grâce à qui, si je suis heureux de partir au bureau le matin, je le suis tout autant d'en revenir le soir.

Chapitre 1

Introduction et contexte applicatif : quelques problématiques rencontrées en épidémiologie

1.1 L'épidémiologie à l'heure des données de grande dimension

1.1.1 Contexte et fléau de la dimension

L'épidémiologie est l'étude des facteurs influant sur l'état de santé de populations, c'est-à-dire l'étude des *causes* de cet état de santé. Elle s'appuie sur des analyses statistiques qui étudient en premier lieu les niveaux d'association entre variables, définis en termes de corrélation ou d'autres mesures telles que l'odds-ratio. Cet état de santé est caractérisé par de multiples composantes : survenue d'une maladie ou d'un accident de la circulation, gravité d'une lésion suite à un accident, etc.. Ces composantes sont typiquement multifactorielles, au sens où elles sont associées à de nombreux facteurs. Le plus souvent, les analyses classiques reposent alors sur des modèles de régression multivariée, recherchant les associations conditionnelles entre la variable d'intérêt, Y, qui décrit une composante particulière de l'état de santé, et un vecteur de covariables ou facteurs de risque, $\mathbf{x} \in \mathbb{R}^p$, $p \geq 1$, décrivant les causes possibles de Y. Ces modèles peuvent ensuite être utilisés, par exemple pour prédire l'état de santé futur des individus. On parle alors de modèles pronostiques. Ils constituent la pierre angulaire de la médecine personnalisée [Hamburg and Collins, 2010]. Un des premiers modèles de ce type, l'équation *de Framingham* publiée en 1976, avait pour objectif de prédire le « risque individuel » de développer une pathologie cardiaque [Kannel et al., 1976]. Des modifications de ce modèle original sont depuis couramment utilisées en clinique afin d'aider à la prise de décision concernant la prévention et les stratégies thérapeutiques. Depuis la fin des années 1980, des modèles pronostiques ont été développés pour prédire le risque de cancer du sein [Gail et al., 1989], puis différents autres types de cancer [Colditz et al., 2000], ou encore le risque de rechute après un premier cancer [Buyse et al., 2006]. Diverses équipes autour de moi ont travaillé, travaillent ou envisagent de travailler à l'élaboration de modèles pronostiques, notamment dans le cas du cancer du sein : l'équipe INSERM dirigée par Françoise Clavel-Chapelon à Villejuif, l'équipe du centre Léon Bérard de David Cox ou encore Pietro Ferrari au Centre International de Recherche

sur le Cancer (CIRC) de l'OMS à Lyon.

L'avènement des données génomiques, protéomiques, métabolomiques, mais aussi celles issues de l'imagerie médicale, ou décrivant l'historique des prescriptions médicamenteuses, ouvre de nouvelles perspectives. Plusieurs modèles ont ainsi été développés, tentant de tirer profit de ces nouvelles sources d'information [McCarthy et al., 2015]. Cependant, ces données posent également de nouvelles questions d'un point de vue méthodologique. D'une part, du point de vue de la qualité de l'estimation, la plupart des procédures statistiques classiques souffrent du *fléau de la dimension* (voir à ce sujet le chapitre introductif du livre de [Giraud, 2014]). Les modèles de régression paramétriques par exemple ont des performances prédictives détériorées lorsqu'ils sont estimés à partir d'un grand nombre de covariables. Or ces performances prédictives sont cruciales dans le cas des modèles pronostiques notamment. D'autre part, du point de vue de l'interprétation, on cherche à travers ces modèles à déterminer quelles covariables sont effectivement associées à la variable d'intérêt, par exemple pour mieux comprendre les mécanismes biologiques en jeu. L'identification des variables pertinentes est cependant d'autant plus difficile que le nombre de variables « candidates » est grand. Ainsi, les données de grande dimension disponibles aujourd'hui posent naturellement la question de la sélection des variables pertinentes, tant pour l'interprétation des modèles obtenus que pour leur garantir de bonnes performances prédictives.

Le problème de la sélection de variables (voire plus généralement de la sélection de modèle) est un des axes de recherche majeurs en statistique. Parmi les procédures classiques de sélection de variables figurent celles qui reposent sur la minimisation de critères pénalisés. Un exemple bien connu est le BIC [Schwarz et al., 1978], pour lequel la consistance en sélection de variable est garantie sous certaines conditions [Kim et al., 2012]. Cependant, ce critère reposant sur la « norme » L_0 des paramètres, il n'est pas convexe et sa résolution numérique est dite combinatoire : il n'existe en général pas d'autres stratégies que celle consistant à calculer le BIC pour l'ensemble des 2^p modèles possibles. Dès que $p \geq 30$, il n'est pas raisonnable de construire les 2^p modèles et on le combine le plus souvent à des heuristiques qui permettent de ne parcourir qu'un sous-ensemble de ces 2^p modèles. Les plus utilisées en épidémiologie et recherche clinique sont les approches « gloutonnes » dites pas-à-pas (*stepwise* en anglais), qui peuvent être ascendantes, descendantes, voire hybrides [Hocking, 1976].

Depuis une vingtaine d'années, la recherche en statistique s'efforce de proposer des critères pénalisés alternatifs, qui soient simples à résoudre numériquement tout en renvoyant des estimateurs présentant de bonnes propriétés statistiques [Candes and Tao, 2007, Tibshirani, 1996, Fan and Li, 2001, Bühlmann and van de Geer, 2011, Giraud, 2014]. Un choix particulier qui a attiré beaucoup d'attention, tant dans la littérature théorique qu'appliquée, est le lasso décrit dans [Tibshirani, 1996]. Il consiste à remplacer la norme L_0 du BIC par son enveloppe convexe sur l'intervalle $[-1, 1]$ [Jojic et al., 2011], à savoir la norme L_1. Une part importante de mes travaux concerne le lasso ou ses dérivés. Le paragraphe suivant le présente brièvement dans le cas du modèle de régression linéaire homoscédastique sur design déterministe, pour simplifier l'exposé. Pour une mise en perspective avec les critères de type BIC un peu plus détaillée, le lecteur peut se référer à l'annexe A.

1.1.2 Le lasso

Pour tout entier $m \geq 1$, notons $[m]$ l'ensemble $\{1, \ldots, m\}$. Nous supposerons disposer d'une matrice déterministe $\mathbf{X} \in \mathbb{R}^{n \times p}$, renfermant les n observations $\mathbf{x}_i$ du vecteur des covariables, pour $i \in [n]$. On notera $X_j \in \mathbb{R}^n$ la j-ème colonne de $\mathbf{X}$, correspondant aux n observations de la j-ème covariable. On suppose disposer par ailleurs d'un échantillon $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ de n observations d'une variable aléatoire d'intérêt, sous le modèle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}. \tag{1.1}$$

On supposera que les composantes du vecteur $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ sont indépendantes et identiquement distribuées (*i.i.d.*), par exemple selon une loi normale $\mathcal{N}(0, \sigma^2)$ avec $\sigma > 0$ fixe mais inconnu. Le vecteur $\boldsymbol{\beta}^* \in \mathbb{R}^p$ renferme les paramètres du modèle à estimer, et décrit l'association entre Y et $\mathbf{x}$. Un estimateur classique $\tilde{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}^*$ est obtenu par la méthode dite des moindres carrés ordinaires (MCO) et est défini par

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Le fléau de la dimension évoqué plus haut peut être illustré ici. Le cadre asymptotique classique, où p est fixe et $n \to \infty$, n'étant pas bien adapté pour le faire, nous supposons que $p = p(n)$ est une fonction croissante de n. Si la matrice de design $\mathbf{X}$ est de rang p (ce qui implique notamment que $p \leq n$), on peut établir l'unicité de la solution $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ dont l'*erreur de prédiction quadratique moyenne* associée est de l'ordre de

$$\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} = \mathcal{O}_{\mathbb{P}}\Big(\frac{p}{n}\Big).$$

Si p est fixe et $n \to \infty$ (qui correspond au cadre asymptotique classique, adapté pour décrire les données où $n \gg p$), ce résultat établit qu'avec probabilité tendant vers 1, l'erreur de prédiction quadratique moyenne tend vers 0 à la vitesse n^{-1}. Cependant, si $p = n^\alpha$, avec $0 < \alpha < 1$, la vitesse de convergence vers 0 de l'erreur de prédiction moyenne est réduite à $n^{-(1-\alpha)}$. Considérons maintenant le cas où $p = n$ et $\mathbf{X} = \mathbf{I}_n$ est la matrice identité d'ordre n. Ce modèle correspond à la version tronquée du modèle de suites gaussiennes [1] : $Y_i = \beta_i^* + \varepsilon_i$, pour $i \in [n]$, avec $\beta_i^* \in \mathbb{R}$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\sigma^2 > 0$. L'estimateur des MCO vaut alors $\tilde{\boldsymbol{\beta}} = \mathbf{Y}$: les espérances β_i^* sont donc chacune estimées par chacune des observations Y_i et

$$\mathbb{E}\left\{\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n}\right\} = \mathbb{E}\left\{\frac{\|\mathbf{Y} - \boldsymbol{\beta}^*\|_2^2}{n}\right\} = \mathbb{E}\left\{\frac{\|\boldsymbol{\varepsilon}\|_2^2}{n}\right\} = \sigma^2.$$

Avec l'estimateur des MCO, l'espérance de l'erreur de prédiction quadratique moyenne ne tend tout simplement pas vers 0 sous ce modèle.

Les approches pénalisées vont permettre d'obtenir des estimateurs affichant de meilleures performances, sous certaines hypothèses, en tirant profit de certaines connaissances a priori. En particulier, dans la plupart des applications, seul un sous-ensemble des covariables est

1. Ce modèle sera dit de suite gaussienne tronquée par la suite.

réellement associé à la variable réponse Y. Ainsi, en notant $J^* = \{j \in [p] : \beta_j^* \neq 0\}$ le support, inconnu, de $\boldsymbol{\beta}^*$ et $p_0 = |J^*|$ le cardinal de J^*, on a typiquement $p_0 \ll p$ et le vecteur $\boldsymbol{\beta}^*$ est alors dit creux ou sparse. Dans de telles situations, les approches pénalisées qui utilisent un terme de pénalité encourageant la sparsité du vecteur solution, comme le lasso, sont particulièrement adaptées. Pour tout $\lambda \geq 0$, les estimateurs lasso sont définis comme solution du problème d'optimisation convexe suivant

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{1.2}$$

où $\|\boldsymbol{\beta}\|_1 = \sum_{j\in[p]} |\beta_j|$ est la norme L_1 du vecteur $\boldsymbol{\beta}$. Comme les critères de type BIC, le critère (1.2) est la somme de deux quantités. La première mesure l'adéquation aux données. La seconde pénalise plus ou moins fortement les vecteurs $\boldsymbol{\beta} \in \mathbb{R}^p$: ces vecteurs sont d'autant plus pénalisés que leur norme L_1 est élevée. En vertu des propriétés géométriques de la norme L_1, les solutions du lasso $\hat{\boldsymbol{\beta}}(\lambda)$ sont typiquement creuses, pour des valeurs assez élevées de $\lambda > 0$. En notant $\hat{J}(\lambda) = \{j \in [p] : \hat{\beta}_j(\lambda) \neq 0\}$ leur support, il a été établi que $\hat{J}(\lambda) = J^*$ avec grande probabilité pour un choix approprié du paramètre de pénalité λ, et ce sous des hypothèses portant sur la matrice de design $\mathbf{X}$, le support J^* de $\boldsymbol{\beta}^*$ et la « force du signal » (mesurée par $\beta_{\min}^* = \min_{j\in J^*} |\beta_j^*|$) [Zhao and Yu, 2006, Zou, 2006, Wainwright, 2009]. Le lasso est alors dit consistant en sélection de variables, ou *sparsistent*. L'hypothèse principale portant sur la matrice de design est celle dite d'irreprésentabilité (*irrepresentability condition*). Pour tout sous-ensemble $J \subseteq [p]$, et toute matrice $\mathbf{U}$ de dimension $n \times p$, notons $\mathbf{U}_J$ la matrice de dimension $n \times |J|$ constituée des colonnes de la matrice $\mathbf{U}$ d'index appartenant à J. Pour toute matrice carrée symétrique $\mathbf{U}$ à valeurs réelles, on désigne par $\Lambda_{\min}(\mathbf{U})$ sa plus petite valeur propre. L'hypothèse d'irrepresentabilité requiert que $\Lambda_{\min}(\mathbf{X}_{J^*}^T \mathbf{X}_{J^*}) > 0$, et

$$\max_{j\notin J^*} \|(\mathbf{X}_{J^*}^T \mathbf{X}_{J^*})^{-1} \mathbf{X}_{J^*}^T X_j\|_1 < 1. \tag{1.3}$$

Autrement dit, la condition d'irreprésentabilité stipule que le modèle restreint à J^* est identifiable et que les colonnes de J^{*c} ne sont pas trop alignées sur celles de J^*, où pour tout sous-ensemble $J \subseteq [p]$, $J^c = [p] \setminus J$ désigne le complémentaire de J. Sous des hypothèses un peu moins restrictives sur la matrice de design $\mathbf{X}$, on peut montrer [Bickel et al., 2009, Dalalyan et al., 2014] que l'erreur de prédiction quadratique moyenne est *oraculaire*, de l'ordre de $\mathcal{O}_{\mathbb{P}}(p_0 \log(p)/n)$. Au terme $\log(p)$ (ainsi qu'aux constantes) près, c'est la vitesse que l'on obtiendrait pour l'estimateur des MCO reposant sur la connaissance a priori du support J^* (voir l'annexe A pour plus de détails).

Ainsi, le lasso affiche, sous certaines hypothèses, de bonnes propriétés statistiques : consistance en sélection de variables, erreur de prédiction oraculaire. Cependant, le problème d'optimisation associé n'admet généralement pas de forme explicite, et sa résolution repose sur des approches numériques. Le problème d'optimisation étant convexe, la complexité algorithmique de ces approches est bien plus faible que dans le cas du BIC par exemple. Elle reste cependant typiquement polynomiale en p et en n. D'autre part, dans certaines situations, la matrice de design est tellement grande que des problèmes de mémoire peuvent survenir lors de la résolution numérique du lasso (on ne peut parfois tout simplement

pas charger la matrice $\mathbf{X}$ en mémoire, sauf à utiliser des mécanismes de type mémoire virtuelle). Des méthodes de présélection ont donc été développées, qui permettent d'éliminer des covariables avant même de résoudre le lasso. Le but est de travailler avec une matrice de design de taille plus faible, de manière à accélérer la résolution du lasso, voire de pouvoir tout simplement charger cette matrice dans la mémoire et résoudre le lasso. Dans [VV4], nous avons développé la première méthode de présélection à bénéficier de la propriété suivante : il est garanti que les variables éliminées par notre approche n'auraient de toute façon pas figuré dans le support de la solution du lasso et l'étape de présélection ne modifie donc pas cette solution du lasso. La présentation de cette approche fait l'objet du chapitre 2.

1.1.3 Pénalités structurées : le fused lasso généralisé

Diverses extensions du lasso ont été proposées dans la littérature pour corriger certains de ses défauts, comme le biais des estimations des composantes non nulles : on peut notamment citer la version OLS-Hybrid du lasso [Efron et al., 2004], le lasso adaptatif de [Zou, 2006], ou encore le lasso relaxé de [Meinshausen, 2007]. Nous renvoyons à l'annexe A pour plus de détails sur ces approches.

D'autres extensions concernent l'utilisation de pénalités structurées [Bach et al., 2012] pour tirer profit d'une structure attendue dans le vecteur $\boldsymbol{\beta}^*$, reflétant une certaine structure au niveau des variables. C'est le cas notamment du fused lasso [Tibshirani et al., 2005]. Il a été initialement proposé dans le modèle de suite gaussienne tronquée ($Y_i = \beta_i^* + \varepsilon_i$, avec $\beta_i^* \in \mathbb{R}$ et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ pour $i \in [n]$) et est spécialement adapté lorsque le signal est constant par morceaux. Un exemple d'application est celui des données CGH en génomique où le signal correspond au logarithme du ratio d'une mesure de la quantité d'ADN le long du génome chez un malade par rapport à un individu sain. En l'absence d'anomalie, le ratio vaut 1 et le signal est donc nul. Lorsqu'une partie du chromosome est amplifiée chez le malade on observe un saut dans le signal, etc. Dans ce type d'application, peuvent être creux non seulement le vecteur $\boldsymbol{\beta}^*$, mais aussi le vecteur des différences successives $\boldsymbol{\Delta}\boldsymbol{\beta}^* = (\beta_2^* - \beta_1^*, \ldots, \beta_p^* - \beta_{p-1}^*)^T \in \mathbb{R}^{p-1}$. Dans ce cadre, le fused lasso consiste à résoudre le problème d'optimisation suivant,

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \boldsymbol{\beta}\|_2^2}{2} + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_1 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{1.4}$$

où λ_1 et λ_2 sont deux paramètres de régularisation et $\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_1 = \sum_{j \geq 2}^p |\beta_j - \beta_{j-1}|$. Par rapport au lasso, le fused lasso pénalise le critère des MCO (ici, dans le modèle de suite gaussienne tronquée) non seulement par la norme L_1 du vecteur de paramètre, mais aussi par la norme L_1 du vecteur des différences successives. Il encourage ainsi les solutions $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ creuses et telles que $\hat{\beta}_j(\lambda_1, \lambda_2) = \hat{\beta}_{j-1}(\lambda_1, \lambda_2)$, c'est-à-dire les solutions creuses et constantes par morceaux. Une illustration est donnée sur la figure 1.1.

Notons d_0 le nombre de composantes non nulles de $\boldsymbol{\Delta}\boldsymbol{\beta}^*$. En se concentrant sur la version du fused lasso pur omettant le terme $\lambda_1 \|\boldsymbol{\beta}^*\|_1$ dans le critère (1.4), il est établi dans [Dalalyan et al., 2014] que l'erreur de prédiction quadratique moyenne est, à des termes logarithmiques près, de l'ordre de d_0/n avec grande probabilité, et pour un choix de λ_2

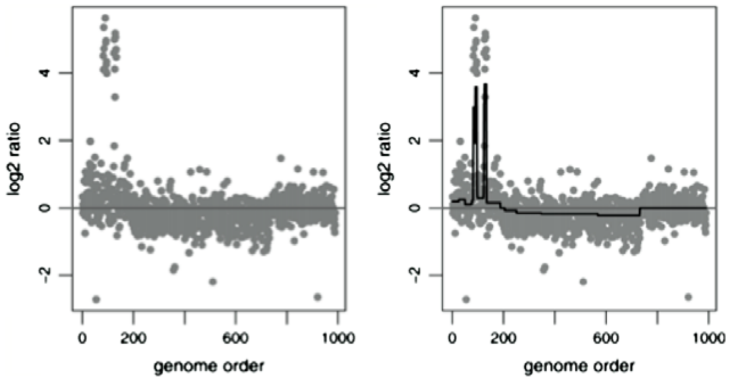


FIGURE 1.1 – Figure reprise de [Tibshirani and Wang, 2008]. Données CGH (à gauche) et estimation du signal par fused lasso (en trait continu sur la figure de droite).

approprié. Là encore, l'utilisation d'une pénalité adaptée à la structure attendue du vecteur des paramètres $\boldsymbol{\beta}^*$ permet d'atteindre la vitesse optimale (à des termes négligeables près).

L'idée du fused lasso a été reprise et généralisée dans le contexte du modèle de régression linéaire (et des modèles linéaires généralisés). Un nombre croissant d'applications fait intervenir des covariables qui sont naturellement organisées en réseau et où il est attendu que des covariables connectées dans le réseau partagent des effets similaires sur la variable réponse considérée. En biologie moléculaire par exemple, les réseaux d'interaction protéines-protéines décrivent les interactions physiques entre protéines. Or des protéines appartenant à une même voie de signalisation, partageant des fonctions proches, peuvent avoir des effets similaires sur la réponse à un traitement ou le développement d'une maladie. Dans les études épidémiologiques de type GWAS (Genome Wide Association Study), on peut également s'attendre à ce que des SNPs (Single Nucleotide Polymorphism) en déséquilibre de liaison ou appartenant à un même gène, etc., partagent des effets similaires sur une pathologie donnée. Si j_1 et j_2 sont deux indices de $[p]$ correspondant à des protéines ou des SNPs connectés dans le réseau, alors on peut s'attendre à ce que $\beta^*_{j_1} = \beta^*_{j_2}$. Le fused lasso généralisé consiste à résoudre le problème d'optimisation suivant,

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j_1 \sim j_2} |\beta_{j_2} - \beta_{j_1}| \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{1.5}$$

dans lequel on pénalise, en plus de la norme L_1 du vecteur de paramètres, les différences $\sum_{j_1 \sim j_2} |\beta_{j_2} - \beta_{j_1}|$ où $j_1 \sim j_2$ signifie que les covariables X_{j_1} et X_{j_2} sont connectées dans le *graphe* décrivant le réseau. En particulier, les termes $|\beta_{j_2} - \beta_{j_1}|$ dans la pénalité encouragent les solutions telles que $\hat{\beta}_{j_1} = \hat{\beta}_{j_2}$. Le fused lasso généralisé encourage donc les vecteurs solution $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ avec une double structure : ces vecteurs auront tendance à être à la fois creux et avec des composantes non nulles égales entre elles pour certaines des covariables connectées dans le graphe.

Dans [VV11], nous nous plaçons dans le cadre asymptotique classique (p fixe et $n \to \infty$) et établissons notamment une propriété oraculaire asymptotique pour une version, dite adaptative, du fused lasso généralisé. Dans ce cadre, notre résultat établit en particulier l'optimalité de la stratégie reposant sur le choix de la clique en tant que graphe décrivant le réseau (la clique est le graphe complet, qui connecte l'ensemble de ses noeuds entre eux ; dans le cadre du fused lasso généralisé, toutes les différences $|\beta_{j_1} - \beta_{j_2}|$, $j_1 < j_2$, sont alors pénalisées). Nous complétons nos résultats théoriques par une étude de simulation approfondie où nous étudions notamment la robustesse du fused lasso généralisé à une mauvaise spécification du graphe par rapport à la structure réelle du vecteur $\boldsymbol{\beta}^*$. Ces résultats empiriques viennent tempérer nos résultats asymptotiques, notamment sur la bonne tenue de l'approche utilisant la clique. Ils vont ainsi dans le sens de ceux obtenus par [Sharpnack et al., 2012] sous le modèle de suite gaussienne tronquée, où $\mathbf{X} = \mathbf{I}_n$ et donc $p = n$ n'est pas fixe.

Outre son intérêt pour les applications dans lesquelles les covariables s'organisent naturellement en réseau, le fused lasso généralisé peut être utilisé lorsque les observations proviennent de différentes *strates*, ou sous-groupes, et que l'on cherche à construire conjointement les modèles correspondant à chacune des strates. Je me suis dernièrement beaucoup intéressé aux données de ce type, qui font l'objet du paragraphe suivant.

1.2 Les données stratifiées

1.2.1 L'exemple des modèles pronostiques pour le cancer du sein

Reprenons l'exemple de la construction d'un modèle pronostique dans le cas du cancer du sein. Les données moléculaires, notamment, ont conduit à la définition de plusieurs sous-types de cancer du sein. Le risque de rechute (ou de décès) après un diagnostic de cancer du sein dépend fondamentalement de ce sous-type de cancer. D'autre part, certains facteurs de risque établis pour le cancer du sein, tels que l'obésité ou le statut ménopausique, ont des effets distincts en fonction du sous-type [Rosner et al., 2013, Tamimi et al., 2012]. On est donc amené à présent à construire des modèles pronostiques pour chacun de ces sous-types. La manière la plus classique de procéder consiste à considérer chaque sous-type isolément (indépendamment) [Munsell et al., 2014, Suzuki et al., 2009, Colditz et al., 2004], ce qui soulève plusieurs problèmes.

Notons $K \geq 1$ le nombre de sous-types considérés. Dans un modèle paramétrique, ou semi-paramétrique comme le modèle de Cox qui est souvent utilisé dans ce contexte [Cox, 1972], le nombre de paramètres à estimer pour construire les K modèles pronostiques correspondant aux K sous-types de cancer du sein est typiquement Kp. Or, même si des hétérogénéités existent entre ces K sous-types, un certain niveau d'homogénéité est attendu : l'effet de certains facteurs peut être identique sur l'ensemble, ou au moins un sous-ensemble, des sous-types. En construisant les K modèles de manière indépendante, on ne peut tirer profit de cette homogénéité. On estime alors un nombre de paramètres inutilement grand, les estimations ont une variance typiquement élevée et finalement les modèles pronostiques ont un pouvoir prédictif modeste (en raison du fléau de la dimension évoqué plus haut). D'autre part, le pouvoir prédictif n'est généralement pas le seul enjeu lorsque

l'on construit un modèle pronostique. Les épidémiologistes s'intéressent également aux variables qui le constituent et aux paramètres qui leur sont associés. Dans le cas d'un modèle pronostique pour plusieurs sous-types de cancer du sein, on s'intéresse en particulier aux différences entre les paramètres correspondant à un même facteur de risque, pour déterminer si son effet varie en fonction du sous-type. Là encore, la stratégie consistant à construire chaque modèle pronostique indépendamment ne permet pas d'interpréter les différences observées puisque les paramètres estimés pour un même facteur sur chacun des sous-types sont différents par construction. Des procédures de test existent [Lunn and McNeil, 1995], mais ne fournissent qu'une réponse partielle en ne permettant de tester que certaines égalités parmi les paramètres (voir le paragraphe suivant).

D'un point de vue général, l'estimation du risque de rechute (ou de décès) pour les K sous-types de cancer du sein peut être vu comme un cas particulier d'apprentissage multi-tâches [Evgeniou and Pontil, 2004, Argyriou et al., 2008], où l'on cherche à estimer une même probabilité conditionnelle dans K strates. L'estimation du risque de survenue de chaque sous-type est un problème différent, faisant intervenir la notion de risques compétitifs [Kalbfleisch and Prentice, 2011, Andersen et al., 2012, Aalen et al., 2008]. Cependant, l'estimation peut être faite sous un modèle de Cox dit stratifié, où les strates correspondent à chacun des sous-types (voir le paragraphe 4.4.2). Ainsi, ces deux exemples illustrent la situation où un facteur de risque catégoriel Z, définissant les strates, revêt un intérêt particulier et peut modifier les effets des autres facteurs sur une variable réponse donnée. Ils décrivent donc la situation classique où l'on cherche à identifier une éventuelle interaction, et à la décrire précisément, le cas échéant. S'agissant dans ce contexte de l'interaction entre une variable catégorielle et un ensemble de covariables, la variable Z est parfois appelée *categorical effect modifier* [Gertheiss and Tutz, 2012, Oelker et al., 2014].

1.2.2 Formulation dans le cas du modèle de régression linéaire

Pour simplifier, considérons à nouveau le cas du modèle linéaire homoscédastique sur design déterministe. Les données de l'échantillon de taille $n \geq 1$ dont on dispose correspondent aux observations des variables $(Y_i, \mathbf{x}_i, Z_i)$, $i \in [n]$, où $Y_i \in \mathbb{R}$ est la variable d'intérêt, $\mathbf{x}_i \in \mathbb{R}^p$ le vecteur des covariables, et $Z_i \in [K]$ la variable catégorielle décrivant la strate d'appartenance de l'observation i. Soit $n_k = \sum_{i\in[n]} \mathbb{I}(Z_i = k)$, le nombre d'observations de la strate k, si bien que $n = \sum_{k\in[K]} n_k$. Pour tout $k \in [K]$, on définit $\mathbf{Y}^{(k)} = (y_1^{(k)}, \ldots, y_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$ le vecteur de variables réponse et $\mathbf{X}^{(k)} = ({\mathbf{x}_1^{(k)}}^T, \ldots, {\mathbf{x}_{n_k}^{(k)}}^T)^T \in \mathbb{R}^{n_k \times p}$ la matrice de design correspondant aux observations de la strate k, c'est-à-dire aux observations $i \in [n]$ telles que $Z_i = k$. On définit par ailleurs $\boldsymbol{\varepsilon}^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$ le vecteur des résidus dans cette strate, dont on supposera qu'il vérifie $\mathbb{E}\boldsymbol{\varepsilon}^{(k)} = \mathbf{0}_{n_k}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}^{(k)}) = \sigma^2 \mathbf{I}_{n_k}$, avec $\sigma^2 > 0$ inconnu. Travailler sous l'hypothèse du modèle linéaire revient ici à considérer que les vecteurs $\mathbf{Y}^{(k)}$ sont liés aux matrices de design $\mathbf{X}^{(k)}$ à travers les K modèles de régression linéaire suivants :

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)} \boldsymbol{\beta}_k^* + \boldsymbol{\varepsilon}^{(k)} \quad \text{pour tout } k \in [K], \tag{1.6}$$

où les vecteurs de paramètres $\boldsymbol{\beta}_k^* \in \mathbb{R}^p$ sont fixes mais inconnus. Ces K modèles décrivent chacun l'association entre Y et $\mathbf{x}$ sur une des K strates. Ils reviennent à supposer que

$$Y = \Big[\sum_{k \in [K]} \mathbb{I}(Z = k) \mathbf{x}^T \boldsymbol{\beta}_k^* \Big] + \varepsilon. \tag{1.7}$$

L'approche naïve estime les K modèles (1.6) indépendamment et, comme évoqué dans le paragraphe précédent, estime donc Kp paramètres (cette complexité peut être ramenée à $\sum_{k \in [K]} \|\boldsymbol{\beta}_k^*\|_0$ estimant chacun des K modèles par des méthodes adaptées si les vecteurs $\boldsymbol{\beta}_k^*$ sont creux). D'autre part, elle renvoie des estimateurs tels que pour tout $j \in [p]$, pour tout $(k, \ell) \in [K]^2$ avec $k \neq \ell$, on a typiquement $\hat{\beta}_{k,j} \neq \hat{\beta}_{\ell,j}$: les différences observées ne peuvent donc pas s'interpréter en termes d'effet de la variable Z sur le lien entre Y et $\mathbf{x}$. On pourrait bien sûr imaginer comparer le modèle imposant la contrainte $\hat{\beta}_{k,j} = \hat{\beta}_{\ell,j}$ et le modèle sans cette contrainte pour tester l'hypothèse $\beta_{k,j}^* \neq \beta_{\ell,j}^*$. Cependant, le nombre total de modèles à considérer pour déterminer, pour tout $j \in [p]$, les couples $(k_1, k_2) \in [K]^2$ tels que $\beta_{k_1,j}^* \neq \beta_{k_2,j}^*$ vaut $(B_K)^p$, où B_K est le nombre de Bell pour K groupes [Bell, 1934]. Dans le cas de 5 groupes et p variables par exemple, on obtient 52^p modèles possibles, si bien que cette procédure est généralement impossible à appliquer en pratique.

Une autre stratégie classique en épidémiologie consiste à sélectionner une strate de référence ℓ, *a priori*, puis à décomposer les paramètres des modèles (1.6) selon l'équation $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_\ell^* + \boldsymbol{\delta}_k^*$, pour tout $k \in [K]$, avec $\boldsymbol{\delta}_\ell^* = \mathbf{0}_p$. Cette stratégie revient à coder la classe d'appartenance par $K-1$ *dummy variables*, c'est-à-dire $K-1$ variables indicatrices $\mathbb{I}(Z = k)$, pour $k \in [K] \setminus \ell$, et à considérer le modèle suivant :

$$Y = \mathbf{x}^T \boldsymbol{\beta}_\ell^* + \sum_{k \neq \ell} (\mathbf{x} \cdot \mathbb{I}(Z = k))^T \boldsymbol{\delta}_k^* + \varepsilon. \tag{1.8}$$

Il correspond à une reparamétrisation du modèle (1.7) et donc des modèles (1.6). Chaque vecteur $\boldsymbol{\delta}_k^*$ renferme ici les différences des effets, pour les p covariables, entre la strate k et la strate de référence ℓ. Une fois ces paramètres estimés, on peut procéder à des tests de significativité, soit pour tester la nullité de chaque composante $\delta_{k,j}^*$, soit pour tester la nullité globale des $\delta_{k,j}^*$ pour tout $k \neq \ell$ (et pour un $j \in [p]$ fixé).

Cette stratégie présente deux défauts principaux. Premièrement, le choix de la strate de référence est arbitraire alors que la précision de l'estimation dépend étroitement de ce choix. Le nombre de paramètres non nuls du modèle reparamétré suite au choix ℓ de la strate de référence est $\|\boldsymbol{\beta}_\ell^*\|_0 + \sum_{k \neq \ell} \|\boldsymbol{\delta}_k^*\|_0$: il dépend donc de ℓ. Considérons la situation où $\beta_{k,j}^* \neq 0$ pour tout $(k, j) \in [K] \times [p]$, $\boldsymbol{\beta}_2^* = \ldots = \boldsymbol{\beta}_K^*$ et, pour tout $j \in [p]$, $\beta_{1,j}^* \neq \beta_{2,j}^*$. Alors le choix $\ell = 1$ pour la strate de référence est associé à une dimension Kp, alors que tout autre choix $\ell \neq 1$ est associé à une dimension $2p < Kp$. Ainsi, dans ce cas, si l'on fait le choix $\ell = 1$ pour la strate de référence, les estimateurs seront moins précis, la puissance pour détecter les composantes $\delta_{k,j}^* \neq 0$ sera plus faible, et le pouvoir prédictif du modèle obtenu sera dégradé, par rapport à tout autre choix de la strate de référence.

Le deuxième défaut de cette stratégie est qu'elle ne fournit qu'une réponse partielle à la question du rôle de la variable Z sur l'association entre $\mathbf{x}$ et Y. Sous le modèle de régression linéaire (1.6), répondre à cette question revient à identifier pour tout $j \in [p]$ les couples

$(k_1, k_2) \in [K]^2$ tels que $\beta^*_{k_1,j} = \beta^*_{k_2,j}$. La stratégie décrite ici ne permet que de tester l'égalité des composantes $\beta^*_{\ell,j}$ et $\beta^*_{k,j}$, pour tout $j \in [p]$ et $k \neq \ell$, mais pas celle des composantes $\beta^*_{k_1,j}$ et $\beta^*_{k_2,j}$ pour k_1 et k_2 différents de ℓ.

Je me suis intéressé à des approches pénalisées permettant d'aborder la problématique des données stratifiées et, plus généralement, le cadre de l'estimation conjointe de K vecteurs de paramètres $\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_K$, sous l'hypothèse d'un certain niveau d'homogénéité entre ces vecteurs. Sous cette hypothèse, on s'attend à ce que des composantes d'une même ligne de la matrice $\mathbf{B}^* = (\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_K)$ (correspondant aux effets d'une même variable dans différentes strates) soient égales. Le principe général des approches que j'ai considérées, et qui seront décrites au chapitre 4, est d'utiliser des pénalités adaptées à cette structure attendue dans la matrice $\mathbf{B}^* = (\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_K)$. Nous montrons en particulier que l'approche proposée par [Gertheiss and Tutz, 2012] correspond à une version du fused lasso généralisé, pour un choix particulier du graphe utilisé dans la pénalité. Un corollaire du résultat obtenu dans [VV11] permet d'établir l'optimalité de la version adaptative de cette approche dans le cadre asymptotique classique. Dans [VV7], nous étendons cette approche au cas des modèles non-linéaires à effets mixtes, qui sont notamment utilisés en pharmacocinétique. Dans [VV8], nous développons une nouvelle approche, AutoRefLasso, qui corrige le premier défaut de la stratégie reposant sur un choix a priori de la strate de référence décrite ci-dessus. Nous étudions ses propriétés en matière de sélection de variables dans un cadre non-asymptotique, et montrons sa supériorité par rapport à la version pénalisée par la norme L_1 de la stratégie reposant sur un choix a priori de la strate de référence, RefLasso. Nous montrons également qu'AutoRefLasso peut se réécrire comme un simple lasso sur une transformation des données originales. Ainsi, premièrement, le coût de sa résolution numérique est peu supérieur à celui de RefLasso (pour de meilleures garanties théoriques). Deuxièmement, AutoRefLasso est directement implémentable sous une variété de modèles (linéaire, logistique, logistique conditionnelle, de Poisson, de Cox, etc.) puisqu'il suffit de disposer d'un algorithme résolvant le lasso sous le modèle considéré.

1.2.3 Extensions

Certains de mes projets concernent diverses extensions des approches présentées dans le paragraphe précédent dans le cadre des modèles de régression. Ces projets sont motivés par des applications concrètes en épidémiologie.

Une des thématiques principales de l'UMRESTTE, mon laboratoire de rattachement, est l'épidémiologie du risque routier. Dans le contexte des accidents de la circulation, la sécurité secondaire s'intéresse aux lésions subies par les victimes de ces accidents. Lorsque les secours arrivent sur les lieux de l'accident, il est important pour eux d'évaluer le plus précisément possible la gravité des lésions subies par chacune des victimes afin de les orienter vers des services hospitaliers adaptés. Or les traumatismes subis par les victimes étant le plus souvent fermés (par opposition aux traumatismes subis par les personnes agressées à l'arme blanche par exemple), le diagnostic de certaines lésions est délicat, comme celles touchant les organes internes. Afin d'aider au diagnostic de ces lésions, on peut chercher à prédire leur présence en fonction notamment des autres lésions subies. Une manière d'aborder cette question est de décrire les associations entre lésions chez les victimes d'accident de la circulation. Or ces

associations peuvent varier en fonction des circonstances de l'accident, et notamment du type d'usager (automobiliste, piéton, cycliste, etc.). Ainsi, pour étudier les associations entre lésions chez les victimes d'accident de la circulation, il semble assez naturel de considérer la population des victimes comme un ensemble de strates définies par les circonstances de l'accident ; voir le paragraphe 5.3. Je me suis initialement intéressé à l'étude des associations parmi un ensemble de variables binaires sur les données du CepiDC. Celles-ci recensent l'ensemble des certificats de décès survenus en France, sur lesquels sont indiquées les causes du décès. L'étude des associations entre ces causes, que nous avons initiée dans [VV9], peut conforter les connaissances actuelles sur les séquences causales conduisant au décès, voire les compléter en en suggérant de nouvelles. Là encore, ces associations varient typiquement en fonction de l'âge et du sexe des individus et il paraît naturel de considérer des strates définies en croisant le sexe et la classe d'âge lorsqu'on étudie ces associations. Ainsi, un de mes projet concerne les extensions des approches évoquées au paragraphe précédent pour estimer simultanément plusieurs modèles graphiques, décrivant chacun les relations d'indépendances conditionnelles parmi un ensemble de variables, sur une strate particulière. Il sera présenté au chapitre 5.

En reprenant l'étude des facteurs de risque des différents sous-types de cancer du sein évoquée au paragraphe précédent, deux designs d'étude sont le plus souvent utilisés : les études de cohorte et les études cas/témoins. Dans les étude de cohorte, des individus sains à l'inclusion dans l'étude sont suivis sur une période de temps donnée et le temps de survenue du cancer (ainsi que le sous-type) est relevé au cours du suivi, le cas échéant. Les différents sous-types de cancer peuvent être considérés comme des risques compétitifs, qui peuvent chacun être modélisés par un modèle de Cox [Cox, 1972]. Comme nous l'avons évoqué plus haut, l'estimation de ces différents risques, en fonction des covariables, peut être effectuée à partir d'un modèle de Cox stratifié. L'extension d'AutoRefLasso dans ce cadre est un de mes projets, présenté au paragraphe 4.4.2.

Dans le cas des études cas/témoins prenant en compte le sous-type de cancer du sein, on dispose de n_0 patients sans cancer du sein, de n_1 patients ayant un cancer du sein de type 1, n_2 de type 2, ..., n_K de type K. C'est notamment le design de l'étude prévue dans un projet financé par l'INCa et porté par Sabina Rinaldi du CIRC (Centre International de Recherche sur le Cancer, OMS), auquel je participe. Il vise à étudier le lien entre l'obésité et le risque des différents sous-types de cancer du sein, notamment à travers des variables mesurant le métabolisme. Un modèle d'analyse classique est le modèle de régression logistique polytomique, qui a la forme suivante :

$$\log\left(\frac{\mathbb{P}(Y=k)}{\mathbb{P}(Y=0)}\right) = \mathbf{x}^T\boldsymbol{\beta}_k^*, \quad \text{pour tout } k \in [K],$$

où Y désigne le type de cancer du sein ($Y=0$ pour les patients sans cancer du sein), $\mathbf{x} \in \mathbb{R}^p$ est le vecteur de covariables et $\boldsymbol{\beta}_k^* = (\beta_{k,1}^*, \ldots, \beta_{k,p}^*) \in \mathbb{R}^p$ avec $\beta_{k,j}^*$ le paramètre associé à la covariable j pour le k-ème sous-type de cancer du sein. Ici, on n'est pas à proprement parlé face à des données stratifiées, ni même à un problème d'apprentissage multi-tâches, mais la question est une nouvelle fois celle de l'estimation de K vecteurs $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$, parmi lesquels une certaine homogénéité est attendue. En particulier, on est également intéressé ici par la détermination des paires $(k_1, k_2) \in [K]^2$ telles que $\beta_{k_1,j}^* = \beta_{k_2,j}^*$ pour $j \in [p]$ fixé. Un de

mes objectifs dans ce projet sera d'étudier l'intérêt des approches que j'ai étudiées dans le cadre des données stratifiées, pour la détection des hétérogénéités parmi les K vecteurs d'un modèle de régression polytomique. Pour être complet, notons que l'on procède le plus souvent à des appariements dans les études cas/témoins et les échantillons de cas et de témoins ne sont alors plus indépendants : l'extension à ce type de donnée pourra également être considérée.

1.3 Problématiques plus spécifiques à l'épidémiologie

Même si j'ai été sensibilisé aux problématiques décrites ci-dessus à travers des applications en épidémiologie, on les retrouve dans de nombreux autres domaines d'application des statistiques. Je me suis intéressé à deux autres types de problèmes, plus spécifiques à l'épidémiologie et la recherche clinique, et qui sont présentés dans les deux paragraphes suivants.

1.3.1 Evaluation des modèles pronostiques en présence de données censurées

Le premier concerne une nouvelle fois les modèles pronostiques, et plus particulièrement leur évaluation. S'agissant de modèles visant à prédire l'état de santé futur, le design privilégié pour construire puis évaluer ces modèles est celui des études de *cohorte prospective*. Dans celles-ci, on inclut un échantillon représentatif des individus sains (qui n'ont pas encore expérimenté l'évènement d'intérêt) de la population cible, qui est ensuite suivi sur une certaine période de temps au cours de laquelle on relève l'instant $\mathcal{T}$ de survenue de l'évènement d'intérêt pour chaque patient. Cependant, les patients inclus ne développeront généralement pas tous la pathologie pendant l'étude, certains patients pouvant par ailleurs être *perdus de vue* avant la fin de l'étude (et possiblement avant d'avoir développé la pathologie). Pour ces individus, on ne dispose que d'une borne inférieure sur $\mathcal{T}$. Ce phénomène est celui de la *censure à droite*, et il est typique de l'analyse de survie dont la construction et l'évaluation des modèles pronostiques sont deux exemples.

Même si d'autres critères existent, deux grandes familles de critères prédominent pour évaluer un modèle pronostique [Gail and Pfeiffer, 2005] : les critères évaluant la calibration, et ceux évaluant le pouvoir discriminant. La calibration mesure l'adéquation du modèle pronostique, et évalue s'il prédit correctement le nombre d'évènements dans des sous-groupes de la population. Or la présence de perdus de vue avant le temps t_0 fait que le nombre d'évènements que l'on aurait observé avant t_0 si tous les patients avaient été au moins suivis jusqu'en t_0 n'est pas connu. Le pouvoir discriminant d'un modèle pronostique mesure quant à lui la capacité du modèle à distinguer les patients qui développeront la maladie avant t_0 de ceux qui ne l'auront toujours pas développée en t_0. La plupart des critères qui l'évaluent sont ainsi des mesures de la distance entre deux distributions : celle des valeurs du modèle pronostique chez les individus qui développeront la maladie avant t_0 et celle des valeurs du modèle pronostique chez les individus qui ne développeront pas la maladie avant t_0. Or on ne sait pas si les individus perdus de vue avant le temps t_0 auraient ou non développé la maladie avant t_0. Ainsi la présence de perdus de vue avant le temps t_0 rend nécessaire

le développement d'estimateurs adaptés, auquel j'ai participé dans [VV14, VV12], pour évaluer sans biais la calibration et le pouvoir discriminant d'un modèle pronostique donné. Nous avons également rédigé un chapitre d'ouvrage présentant une revue de la littérature sur l'évaluation du pouvoir discriminant des modèles pronostiques [VV2], et j'ai co-organisé un atelier INSERM sur ce thème.

Par souci de concision, j'ai cependant décidé de ne pas présenter mes travaux sur cette thématique dans ce document, pas plus que ceux sur la thématique connexe de l'évaluation des tests diagnostiques [VV10, VV3, VV5].

J'ai préféré me concentrer sur ceux que j'ai récemment initiés autour de la causalité, et qui sont introduits dans le paragraphe suivant.

1.3.2 Causalité et effets d'une cause établie

Une problématique à laquelle je me suis intéressé dernièrement est intrinsèque aux objectifs de l'épidémiologie, qui vise à étudier les causes d'un état de santé, et non pas simplement les facteurs qui lui sont associés. Les analyses statistiques classiques qui estiment des mesures d'associations (odds-ratio ajusté, etc.) ne sont donc, en principe, qu'une première étape.

Par exemple, en matière de sécurité routière, la période récente a été marquée par le déploiement des radars automatisés (Contrôle Sanction Automatisé, CSA) durant l'année 2003. Cette mesure s'est accompagnée d'une large diminution des vitesses pratiquées, principalement chez les automobilistes, d'une réduction du nombre d'accidents mortels et en particulier du nombre de décès suite à un traumatisme crânien. J'ai été sollicité par Thomas Lieutaud (Médecin anesthéstiste, UMRESTTE), Blandine Gadegbeku (IR, IFSTTAR, UMRESTTE) et Amina N'diaye (IR, IFSTTAR, UMRESTTE), pour étudier l'évolution de l'épidémiologie des traumatismes crâniens chez les victimes d'accident de la circulation sur les périodes 1996-2001 (avant le CSA) et 2003-2008 (après le CSA). Dans [VV6], nous nous appuyons sur les données du Registre du Rhône et montrons en particulier que la diminution du nombre de décès suite à un traumatisme crânien (-58%) est plus forte que la baisse du nombre de victimes d'un traumatisme crânien dans un accident de la circulation (-42%), cette dernière étant elle-même plus forte que la baisse du nombre d'accidents corporels (-25%). Nous montrons également que ces baisses concernent principalement les automobilistes (chez qui la baisse des vitesses pratiquées suite au CSA est la plus nette). Après ajustement sur différents facteurs mesurant notamment la gravité des lésions subies, on observe un effet protecteur de la période 2003-2008 sur le risque de décès chez les victimes d'un traumatisme crânien (OR ajusté de 0.52, IC à 95% : [0.41, 0.67]), suggérant une meilleure prise en charge de ces victimes dans la période récente. Ainsi, la diminution de 58% du nombre de décès observés suite à un traumatisme crânien chez les victimes d'accident de la circulation entre les deux périodes considérées semble s'expliquer par trois phénomènes principaux : une meilleure prise en charge des victimes, notamment pour les lésions modérées à sévères, une moindre sévérité des accidents corporels et enfin la diminution du nombre de ces accidents. Intuitivement, ces deux derniers phénomènes peuvent au moins en partie être attribués à la baisse des vitesses de circulation observée à la suite du CSA. Cependant, les seules mesures d'association entre la variable binaire décrivant la

période de l'accident (avant ou après 2003) et, par exemple, la sévérité des traumatismes crâniens suite à un accident de la circulation ne suffisent pas à établir le lien causal entre le CSA et cette diminution. Le fait que les associations observées soient plus fortes chez les automobilistes est un argument en faveur de ce lien causal, mais il ne peut être considéré comme suffisant.

Plus généralement, la simple corrélation avec l'état de santé n'est pas suffisante pour qu'un facteur de risque soit qualifié de cause de cet état. En épidémiologie, les critères de Bradford Hill [Hill, 1965], quoique critiquables, ont été proposés pour établir le lien causal entre un facteur de risque et un état de santé : plausibilité, relation dose-effet, reproductibilité, temporalité, spécificité, etc. Pour certains évènements, leurs causes, ou en tout cas certaines d'entre elles, sont ainsi considérées comme établies dans la littérature : le tabac pour le cancer du poumon, plus récemment la consommation de viande rouge pour le cancer, etc. Pour une cause établie, une mesure d'importance en épidémiologie est son risque attribuable, ou fraction attribuable, qui quantifie la proportion des cas de maladie due, ou attribuable, à cette cause [Rothman et al., 2008]. Régulièrement, le CIRC met par exemple à jour les risques attribuables de cancer pour différents facteurs de risque causaux [IARC, 2001]. Dans le domaine de la sécurité routière, une cause bien établie des accidents, et notamment des accidents mortels, est la vitesse. Lors de mon arrivée à l'UMRESTTE, j'ai été sollicité par Bernard Laumon (DR IFSTTAR), alors directeur de l'UMRESTTE, pour étendre les équations de Nilsson [Nilsson, 2004], qui forment un modèle bien connu en sécurité routière. Un des résultats marquants de ces modèles peut se résumer ainsi. Soit t_0 et t_1 deux temps distincts, et pour $j \in \{0, 1\}$, soit v_j et d_j la vitesse moyenne et le nombre d'accidents mortels observés sur un réseau routier donné au temps t_j. Alors on a la relation suivante $d_1/d_0 = (v_1/v_0)^4$. Ce modèle simple a été validé sur un grand nombre d'études (voir par exemple la méta-analyse de [Elvik et al., 2005] portant sur 98 études). L'idée originale de notre travail était de relier le nombre d'accidents mortels non pas à la vitesse moyenne, mais à la distribution complète des vitesses, en supposant une relation polynomiale entre la vitesse d'un groupe de véhicules et leur risque d'être impliqué dans un accident mortel. Dans [VV13], nous avons utilisé les données de vitesse et d'accidentologie collectées au niveau national par l'Organisme National Inter-ministériel de Sécurité Routière (ONISR) en nous focalisant sur les données de jour relatives aux routes départementales et nationales, qui concentrent la part principale du trafic et des accidents mortels. Nous avons construit un modèle qui, malgré sa simplicité (on ne considère en somme que la vitesse comme facteur prédictif du nombre d'accidents mortels), était en bonne adéquation avec nos données. Nous avons ensuite utilisé ce modèle pour estimer les fractions des accidents mortels attribuables à différents types d'excès de vitesse. Par exemple, le nombre d'accidents mortels attribuable aux excès de vitesse compris entre 10 et 20 km/h au dessus de la limite autorisée était estimé en comparant les nombres d'accidents mortels prédits par notre modèle dans la situation observée sur nos données et dans la situation « contrefactuelle » où les conducteurs circulant entre 10 et 20 km/h au-dessus des limites auraient circulé à la vitesse réglementaire. Nos résultats sont en grande partie cohérents avec ceux obtenus via les équations de Nilsson. Ils suggèrent que sur les routes départementales, la fraction des accidents mortels attribuables aux « grands » excès de vitesse (>20 km/h au-dessus de la limite autorisée) est passée de 25% à 6% sur la période 2001-2010, celle des excès modérés (entre 10 et 20 km/h au-dessus

de la limite) est passée de 13% à 9%, alors que la fraction attribuable aux petits excès de vitesse (<10 km/h au-dessus de la limite) est passée de 7% à 13%. Nous avons par ailleurs observé des tendances analogues sur les routes nationales. A noter que ces résultats reflètent surtout le fait que la fréquence des grands excès de vitesse a beaucoup diminué suite au déploiement des radars automatisés en 2003, alors que celle des petits excès de vitesse est restée relativement stable. En toute rigueur, ils sont aussi à considérer avec précaution puisqu'aucun ajustement n'était possible sur des facteurs tels que l'alcool, l'utilisation du téléphone portable au volant, etc.

Dans les situations caractérisées par la présence d'un facteur de risque intermédiaire ou médiateur, on peut par ailleurs chercher à décomposer l'effet d'une cause, en un effet direct et un effet indirect, médié par ce médiateur. Par exemple, dans l'étude du rôle du régime alimentaire, ou plus généralement du mode de vie, sur la survenue d'un cancer, le métabolisme peut être considéré comme un médiateur possible. J'ai été sollicité par Pietro Ferrari (CIRC, OMS) pour participer au co-encadrement de la thèse de Nada Assi, qui a pour objectif général l'étude des effets du mode de vie sur le risque de cancer. Nous avons en particulier publié un article dans lequel nous modélisons l'approche dite ≪ meeting-in-the-middle ≫ [Chadeau-Hyam et al., 2011] où trois ensembles de variables sont en jeu : des variables liées au mode de vie (régime alimentaire, variables anthropométriques, etc.), des variables mesurant le métabolisme et une variable indiquant la survenue d'un cancer du foie. L'idée fondamentale du meeting-in-the-middle est que l'effet du mode de vie sur le risque de cancer (ici du foie), est en partie médié par le métabolisme, ce que semblent confirmer nos résultats [VV1].

Ces trois travaux collaboratifs, [VV6, VV13] et [VV1], mettent en jeu des notions classiques en épidémiologie. Elles sont abordées dans les formations en biostatistique auxquelles j'ai pu participer en tant qu'étudiant (ISUP) ou enseignant (ISUP, Paris 5, Lyon 1). Dans ces formations, on insiste sur la distinction entre l'effet marginal et l'effet ajusté d'un facteur de risque, et donc, sur la nécessité d'ajuster sur des facteurs de confusion pour mesurer au mieux l'effet d'un facteur de risque sur une variable d'intérêt (comme par exemple l'ajustement sur la gravité des lésions pour étudier la meilleure prise en charge des patients dans la période récente dans [VV6]). A contrario, on apprend aussi à ne pas ajuster sur un facteur intermédiaire (ou médiateur), au risque de n'estimer que l'effet direct d'un facteur de risque causal, et donc sous-estimer son effet total. Ces règles sont cependant généralement dictées sans réelle justification formelle. Or, un pan de la littérature récente permet de les justifier sous certaines hypothèses, voire de les étendre sous d'autres hypothèses. Il s'agit de la littérature concernant l'*inférence causale sur données observationnelles* (par opposition aux données interventionnelles de l'essai thérapeutique notamment). L'inférence causale fournit en particulier des définitions de l'*effet causal* pour une cause établie, à partir de variables latentes, dites contrefactuelles ou résultats potentiels [Chambaz et al., 2014, Greenland et al., 1999, Pearl, 2000, Pearl, 2009, Robins, 1986, Rubin, 1974, Rothman et al., 2008]. Ces variables représentent la variable d'intérêt que l'on aurait observée si l'on était intervenu pour imposer une certaine valeur à la cause étudiée, recréant ainsi le cadre des données interventionnelles. Le cadre formel développé notamment par Pearl [Pearl, 2000, Pearl, 2009] permet également de préciser les situations où ces effets causaux sont identifiables et estimables à partir des variables observées. Par exemple, sous

des modèles simples (modèles linéaires sans interaction, etc.), cet effet causal se ramène, au moins approximativement, aux mesures d'associations ajustées classiques telles que le coefficient d'un modèle de régression linéaire multiple ou encore le risque relatif ajusté, etc. L'introduction des variables contrefactuelles permet aussi la définition précise des effets directs et indirects en présence de médiateurs (et les conditions sous lesquelles ces quantités sont identifiables à partir des variables observées).

J'ai commencé à m'intéresser à cette littérature au cours des travaux décrits ci-dessus, et surtout depuis le début de la thèse de Marine Dufournet, que je co-encadre avec Jean-Louis Martin (CR, IFSTTAR, UMRESTTE) et Alain Bergeret (PU-PH, UCBL, HCL, UMRESTTE). L'objectif général de cette thèse est de hiérarchiser les facteurs causaux d'accident de la circulation. Une des particularités des données disponibles dans ce contexte est qu'elles ne concernent en général que des conducteurs impliqués dans des accidents (voire des accidents corporels). L'état de nos réflexions quant à l'identifiabilité des effets causaux sur ces données est présenté au chapitre 6, qui introduit également les principes généraux de l'inférence causale.

Première partie

Résultats généraux autour des approches pénalisées

Chapitre 2

SaFE : Safe Feature Elimination pour le lasso

2.1 Rappels concernant le lasso

On se place dans le cadre du modèle de régression introduit en (1.1), et on considère le problème d'optimisation associé au lasso, à savoir

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p. \tag{2.1}$$

De nombreux algorithmes ont été développés pour résoudre ce problème d'optimisation. Citons par exemple ceux de [Efron et al., 2004, Kim et al., 2007, Park and Hastie, 2007, Donoho and Tsaig, 2008, Friedman et al., 2007, Friedman et al., 2010, Becker et al., 2011]. Cependant, la complexité des ces algorithmes (lorsqu'elle est connue précisément), croît rapidement avec le nombre de covariables p. Alors que les estimateurs lasso sont particulièrement intéressants en présence de données de grande dimension, les algorithmes disponibles peuvent être relativement lents dans de tels contextes. Le problème est d'autant plus important pour les approches nécessitant la résolution de centaines (voire plus) de problèmes de type lasso, telles que Bolasso de [Bach, 2008, Varoquaux et al., 2012], la *stability selection* de [Meinshausen and Bühlmann, 2010], ou encore les méthodes de sélection de la structure des modèles graphiques gaussiens proposées par [Meinshausen and Bühlmann, 2006], et étendues par la suite au cas de modèles graphiques binaires par [Ravikumar et al., 2010]. D'autre part, dans certaines applications la matrice de design $\mathbf{X}$ est tellement grande qu'on ne peut pas résoudre le lasso en raison de problèmes de mémoire (en particulier lorsqu'on ne peut même pas charger cette matrice en mémoire). Ainsi, un champ de recherche actif concerne le développement de méthodes de présélection, ou *screening*, [Fan and Lv, 2008, Xiang et al., 2014]. Elles visent à éliminer des covariables, ou « features », dans une étape préliminaire, afin de réduire la dimension et résoudre le problème d'optimisation sur une matrice de design réduite.

Ces approches sont généralement rapides du point de vue de leur résolution numérique. Leur principe est d'assigner à chaque covariable un score, par exemple la statistique du test de Student ou du χ^2 pour la comparaison de deux échantillons ([Fan and Lv, 2008, Fan and Lv, 2010] ; voir aussi [Forman, 2003] et ses références). Elles éliminent ensuite les

covariables présentant les scores les plus faibles, sans garantie que ces variables éliminées n'auraient pas sinon appartenu au support de la solution retournée par le lasso.

Dans [VV4], nous proposons une approche de présélection, SaFE (pour Safe Feature Elimination), qui était la première à présenter la propriété suivante : toutes les variables éliminées par SaFE n'auraient de toute façon pas été sélectionnées par le lasso ; depuis, les approches vérifiant cette propriété sont dites *safe* dans la littérature [Xiang et al., 2014, Fercoq et al., 2015]. Plus précisément, supposons que l'on cherche à résoudre le lasso avec la valeur λ du paramètre de pénalité et que toute solution $\hat{\boldsymbol{\beta}}(\lambda)$ de (2.1), inconnue à ce stade, soit creuse, c'est-à-dire $|\hat{J}(\lambda)| < p$, avec $\hat{J}(\lambda) = \{j \in [p] : \hat{\beta}_j(\lambda) \neq 0\}$. Posons $\hat{J}^c(\lambda) = [p] \setminus \hat{J}(\lambda)$. SaFE identifie, avant même de résoudre le lasso, un sous-ensemble $S \subseteq \hat{J}^c(\lambda)$, dont les éléments correspondent à des composantes nulles de toute solution possible $\hat{\boldsymbol{\beta}}(\lambda)$ du lasso. On peut ensuite éliminer « sans risque » les colonnes correspondantes de la matrice $\mathbf{X}$ et résoudre le lasso sur la matrice de design réduite $\mathbf{X}_{S^c}$ pour obtenir $\hat{\boldsymbol{\beta}}_{S^c}(\lambda)$ et en déduire une solution $\hat{\boldsymbol{\beta}}(\lambda)$ que l'on aurait pu obtenir en résolvant le lasso sur la matrice de design complète $\mathbf{X}$.

2.2 Principe général de SaFE

Comme précédemment, notons $\hat{\boldsymbol{\beta}}(\lambda)$ une solution du lasso pour un paramètre de pénalité $\lambda \geq 0$ donné, soit

$$\hat{\boldsymbol{\beta}}(\lambda) \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{2.2}$$

Le problème d'optimisation correspondant sera noté $\mathcal{P}(\lambda)$ par la suite et on introduit $\psi(\lambda)$, la valeur optimale de la fonction objectif de $\mathcal{P}(\lambda)$ atteinte en toute solution $\hat{\boldsymbol{\beta}}(\lambda)$.

Le problème d'optimisation $\mathcal{P}(\lambda)$ est appelé problème primal, $\boldsymbol{\beta} \in \mathbb{R}^p$ la variable primale, et $\hat{\boldsymbol{\beta}}(\lambda)$ un point primal optimal (l'unicité de $\hat{\boldsymbol{\beta}}(\lambda)$ n'étant pas garantie). En notant $\Delta_{\mathbf{X}} = \{\boldsymbol{\theta} \in \mathbb{R}^n : |\boldsymbol{\theta}^T X_j| \leq \lambda, \forall j \in [p]\}$, la formulation duale du lasso (2.2) [Kim et al., 2007] s'écrit

$$\hat{\boldsymbol{\theta}}(\lambda) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Delta_{\mathbf{X}} \subset \mathbb{R}^n} G(\boldsymbol{\theta}) \quad \text{avec } G(\boldsymbol{\theta}) := \frac{1}{2}\|\mathbf{Y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta} + \mathbf{Y}\|_2^2. \tag{2.3}$$

En notant $\prod_{\mathcal{C}}$ la projection sur un ensemble convexe $\mathcal{C}$, il vient $\hat{\boldsymbol{\theta}} = \prod_{\Delta_{\mathbf{X}}}(-\mathbf{Y})$ ce qui garantit l'unicité de la solution $\hat{\boldsymbol{\theta}}(\lambda)$. On note $\mathcal{D}(\lambda)$ le problème d'optimisation dual. Celui-ci est un problème d'optimisation convexe sur la variable duale $\boldsymbol{\theta} \in \mathbb{R}^n$. Un point $\boldsymbol{\theta}$ est dit dual faisable s'il appartient à l'ensemble $\Delta_{\mathbf{X}}$, qu'on appelle l'ensemble dual faisable. Le lasso (2.2) vérifie la propriété de dualité forte, si bien que la valeur optimale de $\mathcal{D}(\lambda)$ atteint $\psi(\lambda)$ au point dual optimal $\hat{\boldsymbol{\theta}}(\lambda)$, solution de (2.3). D'autre part, à l'optimum, on a $\hat{\boldsymbol{\theta}}(\lambda) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda) - \mathbf{Y}$.

Nous avons recours au problème dual $\mathcal{D}(\lambda)$ en raison d'une propriété fondamentale, sur laquelle repose notre approche (et toutes les approches « safe » qui ont suivi). Supposons que $\hat{\boldsymbol{\beta}}(\lambda)$ soit creux. Alors la connaissance de $\hat{\boldsymbol{\theta}}(\lambda)$ nous permettrait d'identifier certaines composantes nulles dans $\hat{\boldsymbol{\beta}}(\lambda)$. En effet, les conditions d'optimalité du premier

ordre assurent que

$$\forall j \in [p], X_j^T \hat{\boldsymbol{\theta}}(\lambda) \begin{cases} = -\lambda \text{sign}(\hat{\beta}_j(\lambda)) & \text{si} \quad \hat{\beta}_j(\lambda) \neq 0 \\ \in [-\lambda, \lambda] & \text{si} \quad \hat{\beta}_j(\lambda) = 0, \end{cases} \tag{2.4}$$

où $\text{sign}(x) = 1$ si $x > 0$, -1 si $x < 0$ et 0 si $x = 0$. On en déduit la propriété suivante [Boyd and Vandenberghe, 2004] :

$$|X_j^T \hat{\boldsymbol{\theta}}(\lambda)| < \lambda \Rightarrow \hat{\beta}_j(\lambda) = 0, \tag{2.5}$$

et ce pour toute solution possible $\hat{\boldsymbol{\beta}}(\lambda)$. Ce résultat ne nous permet pas à lui seul d'éliminer des colonnes a priori puisque le point dual optimal $\hat{\boldsymbol{\theta}}(\lambda)$ n'est pas connu. On peut cependant exploiter les implications de (2.5). Plus précisément, notre approche consiste à construire un sous-ensemble de points duaux faisables $\Theta \subset \Delta_{\mathbf{X}} \subset \mathbb{R}^n$, avant de résoudre le lasso, qui vérifie les deux propriétés suivantes :

$$\Theta \text{ contient le point dual optimal : } \hat{\boldsymbol{\theta}}(\lambda) \in \Theta. \tag{2.6}$$

$$\max_{\boldsymbol{\theta} \in \Theta} |X_j^T \boldsymbol{\theta}| < \lambda \text{ pour certaines colonnes } X_j. \tag{2.7}$$

Dès lors que ces deux conditions sont vérifiées, alors il est garanti que $|X_j^T \hat{\boldsymbol{\theta}}(\lambda)| < \lambda$, et donc que $\hat{\beta}_j(\lambda) = 0$: la j-ème colonne X_j peut être éliminée de la matrice $\mathbf{X}$, sans risque.

2.3 Mise en oeuvre

Le plus souvent en pratique, on ne cherche pas à résoudre le lasso pour une seule valeur particulière du paramètre λ, mais plutôt pour une séquence de valeurs, du type $\lambda_{\max} \geq \lambda_1 \geq \cdots \geq \lambda_N$. Par exemple [Bühlmann and van de Geer, 2011] [2.12.1] suggèrent la séquence $\lambda_k = \lambda_{\max} 10^{-\delta k/(N-1)}$, avec $\delta > 0$. La valeur $\lambda_{\max}$ correspond à $\min\{\lambda \geq 0 : \forall \lambda' \geq \lambda, \hat{\boldsymbol{\beta}}(\lambda') = \mathbf{0}_p\}$, c'est-à-dire la plus petite valeur au-delà de laquelle l'unique solution du lasso est le vecteur nul $\mathbf{0}_p$. On montre facilement que $\lambda_{\max} = \|\mathbf{X}^T \mathbf{Y}\|_\infty$. La solution du lasso étant connue pour $\lambda = \lambda_{\max}$, nous pouvons nous placer sans perte de généralité dans le contexte suivant. Etant donnés deux réels $\lambda_0 \geq \lambda \geq 0$, nous supposons que la solution duale optimale $\hat{\boldsymbol{\theta}}(\lambda_0)$ de $\mathcal{D}(\lambda_0)$ et une solution primale optimale $\hat{\boldsymbol{\beta}}(\lambda_0)$ de $\mathcal{P}(\lambda_0)$ sont connues, et que nous cherchons à éliminer des colonnes de $\mathbf{X}$ avant de résoudre le problème $\mathcal{P}(\lambda)$.

Nous décrivons dans ce paragraphe une approche de construction d'un ensemble Θ qui vérifie les hypothèses (2.6) et (2.7). Evidemment, plus cet ensemble Θ est petit, plus la quantité $P(X_j) := \max_{\boldsymbol{\theta} \in \Theta} |X_j^T \boldsymbol{\theta}|$ de la condition (2.7) est petite, et donc plus notre approche est efficace (au sens où elle élimine plus de covariables). Notre objectif est donc de construire, avant de résoudre $\mathcal{P}(\lambda)$ ou $\mathcal{D}(\lambda)$, le plus petit ensemble Θ possible vérifiant la condition (2.6), à savoir $\hat{\boldsymbol{\theta}}(\lambda) \in \Theta$.

D'une part, $\hat{\boldsymbol{\theta}}(\lambda)$ est optimal pour le problème $\mathcal{D}(\lambda)$. La solution $\hat{\boldsymbol{\theta}}(\lambda)$ vérifie donc $G(\hat{\boldsymbol{\theta}}(\lambda)) \geq G(\boldsymbol{\theta})$ pour tout point dual faisable $\boldsymbol{\theta}$ de $\mathcal{D}(\lambda)$. Supposons disposer d'un tel point dual faisable, $\boldsymbol{\theta}_s$, et notons $\Upsilon := G(\boldsymbol{\theta}_s)$. Alors $G(\hat{\boldsymbol{\theta}}(\lambda)) \geq \Upsilon$ et donc $\hat{\boldsymbol{\theta}}(\lambda) \in \Theta_1$ avec $\Theta_1 := \{\boldsymbol{\theta} \in \mathbb{R}^n : G(\boldsymbol{\theta}) \geq \Upsilon\}$. D'autre part, $\hat{\boldsymbol{\theta}}(\lambda)$ étant le point dual optimal, il est dual

faisable et appartient donc à $\Delta_{\mathbf{X}}$. Or l'ensemble Θ_1 peut contenir des points qui ne sont pas dans $\Delta_{\mathbf{X}}$. Nous allons donc chercher à caractériser un ensemble $\Theta_2 \supseteq \Delta_{\mathbf{X}}$ qui contienne l'ensemble des points duaux faisables, et l'on définira finalement $\Theta = \Theta_1 \cap \Theta_2$. Le critère pour éliminer la j-ème colonne avant de résoudre le problème $\mathcal{P}(\lambda)$ sera alors

$$\lambda > \max_{\boldsymbol{\theta} \in \Theta} |X_j^T \boldsymbol{\theta}|.$$

La forme particulière de l'ensemble Θ que nous construisons va en outre nous permettre d'obtenir la forme analytique de $\max_{\boldsymbol{\theta} \in \Theta} |\boldsymbol{\theta}^T X_j|$ et d'évaluer ainsi notre critère simplement (en nous passant notamment d'utiliser un algorithme itératif pour résoudre numériquement $\max_{\boldsymbol{\theta} \in \Theta} |X_j^T \boldsymbol{\theta}|$).

2.3.1 Construction de l'ensemble Θ_1

Pour construire Θ_1, il nous faut trouver un point $\boldsymbol{\theta}_s$ dual faisable pour $\mathcal{D}(\lambda)$, tel que $\Upsilon = G(\boldsymbol{\theta}_s)$ soit la plus élevée possible, de telle sorte que $\Theta_1 = \{\boldsymbol{\theta} \in \mathbb{R}^n : G(\boldsymbol{\theta}) \geq \Upsilon\}$ soit le plus petit possible. Nous disposons du point dual optimal $\hat{\boldsymbol{\theta}}_0$ de $\mathcal{D}(\lambda_0)$. Etant dual optimal pour $\mathcal{D}(\lambda_0)$, il est dual faisable pour $\mathcal{D}(\lambda_0)$ si bien que $\|X^T \hat{\boldsymbol{\theta}}_0\|_\infty \leq \lambda_0$. On peut en fait montrer que $\|X^T \hat{\boldsymbol{\theta}}_0\|_\infty = \lambda_0$ et donc $\hat{\boldsymbol{\theta}}_0$ n'est pas dual faisable pour $\mathcal{D}(\lambda)$ puisque $\lambda < \lambda_0$. On peut par contre construire un point dual faisable $\boldsymbol{\theta}_s$ pour $\mathcal{D}(\lambda)$ en posant $\boldsymbol{\theta}_s = s\hat{\boldsymbol{\theta}}_0$, pour un scalaire $s \geq 0$ assurant que $\|X^T \boldsymbol{\theta}_s\|_\infty \leq \lambda$, c'est-à-dire $|s| \leq \lambda/\lambda_0$. Il ne nous reste plus qu'à optimiser la valeur de ce scalaire s, maximisant la valeur $\Upsilon = G(\boldsymbol{\theta}_s)$. On définit donc Υ à partir du problème d'optimisation suivant :

$$\Upsilon = \max_s \left\{ G(s\hat{\boldsymbol{\theta}}_0) \ : \ |s| \leq \frac{\lambda}{\lambda_0} \right\} = \max_s \left\{ \omega_0 s - \frac{1}{2} s^2 \alpha_0 \ : \ |s| \leq \frac{\lambda}{\lambda_0} \right\},$$

avec $\alpha_0 := \hat{\boldsymbol{\theta}}_0^T \hat{\boldsymbol{\theta}}_0 > 0$ et $\omega_0 := |y^T \hat{\boldsymbol{\theta}}_0|$. On obtient aisément

$$\Upsilon = \frac{\lambda}{\lambda_0}\Big(\omega_0 - \frac{\alpha_0}{2}\frac{\lambda}{\lambda_0}\Big). \tag{2.8}$$

L'ensemble Θ_1 est ensuite simplement défini à partir de cette valeur de Υ,

$$\begin{aligned} \Theta_1 &= \{\boldsymbol{\theta} \in \mathbb{R}^n : G(\boldsymbol{\theta}) \geq \Upsilon\} \\ &= \{\boldsymbol{\theta} \in \mathbb{R}^n : \frac{1}{2}\|\mathbf{Y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta} + \mathbf{Y}\|_2^2 \geq \Upsilon\} \\ &= B(-\mathbf{Y}, R_\Upsilon), \end{aligned}$$

avec $R_\Upsilon = \|\mathbf{Y}\|_2^2 - 2\Upsilon \geq 0$ et $B(\mathbf{x}, R)$ la boule de $\mathbb{R}^n$ de centre $\mathbf{x}$ et de rayon R.

2.3.2 Construction de l'ensemble Θ_2

La construction de l'ensemble Θ_2 repose sur une caractérisation des points duaux faisables pour $\mathcal{D}(\lambda)$. Premièrement, observons que tout point dual faisable $\boldsymbol{\theta}$ pour $\mathcal{D}(\lambda)$ l'est également pour $\mathcal{D}(\lambda_0)$ puisque pour tout $\lambda \leq \lambda_0$, on a

$$\|\mathbf{X}^T \boldsymbol{\theta}\|_\infty \leq \lambda \Rightarrow \|\mathbf{X}^T \boldsymbol{\theta}\|_\infty \leq \lambda_0.$$

D'autre part, on peut caractériser l'ensemble des points duaux faisables pour $\mathcal{D}(\lambda_0)$ grâce à la condition d'optimalité du premier ordre pour les problèmes d'optimisation convexes sous contrainte. D'après celle-ci, pour tout point dual faisable $\boldsymbol{\theta}$ pour $\mathcal{D}(\lambda_0)$, $\nabla G(\hat{\boldsymbol{\theta}}(\lambda_0))^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\lambda_0)) \leq 0$. En d'autres termes,

$$\|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq \lambda_0 \;\Rightarrow\; \nabla G(\hat{\boldsymbol{\theta}}(\lambda_0))^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\lambda_0)) \leq 0.$$

En combinant ces deux résultat et en observant que $\nabla G(\boldsymbol{\theta}) = -(\boldsymbol{\theta} + \mathbf{Y})$, on obtient la caractérisation suivante des points duaux faisables pour $\mathcal{D}(\lambda)$:

$$\|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq \lambda \;\Rightarrow\; (\hat{\boldsymbol{\theta}}(\lambda_0) + \mathbf{Y})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\lambda_0)) \geq 0.$$

Ainsi, le point dual optimal $\hat{\boldsymbol{\theta}}(\lambda)$ est dans le demi-espace

$$\Theta_2 := \{\boldsymbol{\theta} \in \mathbb{R}^n : (\hat{\boldsymbol{\theta}}(\lambda_0) + \mathbf{Y})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\lambda_0)) \geq 0\}. \tag{2.9}$$

2.3.3 Résultat principal

Soit $\Theta = \Theta_1 \cap \Theta_2$, avec Θ_1 et Θ_2 définis aux paragraphes précédents. Notre critère pour déterminer si l'on peut éliminer la j-ème colonne de la matrice de design $\mathbf{X}$ (le j-ème feature) pour le problème $\mathcal{P}(\lambda)$ s'écrit

$$\max_{\boldsymbol{\theta}\in\Theta} |X_j^T\boldsymbol{\theta}| \overset{?}{<} \lambda. \tag{2.10}$$

Une formulation équivalente de la condition (2.10) est

$$\max(P(\Upsilon, X_j), P(\Upsilon, -X_j)) \overset{?}{<} \lambda,$$

où $P(\Upsilon, X_j)$ est la solution du problème d'optimisation sous contrainte suivant :

$$\begin{aligned} P(\Upsilon, X_j) &:= \max_{\boldsymbol{\theta}\in\Theta} X_j^T\boldsymbol{\theta} \\ &= \max_{\boldsymbol{\theta}\in\mathbb{R}^n} X_j^T\boldsymbol{\theta} \,:\, G(\boldsymbol{\theta}) \geq \Upsilon,\, (\hat{\boldsymbol{\theta}}(\lambda_0) + \mathbf{Y})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\lambda_0)) \leq 0 \end{aligned} \tag{2.11}$$

Ce problème d'optimisation convexe est simple à résoudre et admet une forme analytique pour la valeur optimale $P(\Upsilon, X_j)$ (donnée en (2.12) ci-dessous). Finalement, on peut résumer notre approche dans le théorème suivant.

Théorème 2.3.1 *On considère le problème lasso $\mathcal{P}(\lambda)$ en (2.2). Soit $\lambda_0 \geq \lambda$ une valeur du paramètre de pénalité pour laquelle une solution $\hat{\boldsymbol{\beta}}_0 \in \mathbb{R}^p$ est connue. Soit de plus $\hat{\boldsymbol{\theta}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}_0 - \mathbf{Y}$, $g = \hat{\boldsymbol{\theta}}_0 + \mathbf{Y}$, $\alpha_0 = \|\hat{\boldsymbol{\theta}}_0\|_2^2$, $\omega_0 = |\mathbf{Y}^T\hat{\boldsymbol{\theta}}_0|$, $\Upsilon = (\lambda/\lambda_0)[\omega_0 - (\alpha_0\lambda)/(2\lambda_0)]$, $R_\Upsilon = (\|\mathbf{Y}\|_2^2 - 2\Upsilon)^{1/2}$, $\tilde{R}_\Upsilon = [2(G(\hat{\boldsymbol{\theta}}_0) - \Upsilon)]^{1/2}$ et, pour tout $j \in [p]$ $\kappa_j^2 := \|X_j\|_2^2 - \frac{(X_j^T g)^2}{\|g\|_2^2} \geq 0$. Alors la condition*

$$\lambda > \max\Big(P(\Upsilon, X_j), P(\Upsilon, -X_j)\Big),$$

avec

$$P(\Upsilon, X_j) \;=\; \begin{cases} \hat{\boldsymbol{\theta}}_0^T X_j + \kappa_j \tilde{R}_\Upsilon & si \quad \frac{1}{R_\Upsilon} \|g\|_2^2 \|X_j\|_2 \geq X_j^T g, \\ -\mathbf{Y}^T X_j + \|X_j\|_2 R_\Upsilon & sinon \end{cases} \tag{2.12}$$

assure que $\hat{\beta}_j(\lambda) = 0$ *pour toute solution* $\hat{\boldsymbol{\beta}}(\lambda)$ *de* $\mathcal{P}(\lambda)$ *et permet donc d'éliminer sans risque la* j*-ème colonne de* $\mathbf{X}$ *avant de résoudre* $\mathcal{P}(\lambda)$*.*

Considérons une nouvelle fois le cas où le lasso doit être résolu pour une séquence $\lambda_{\max} > \lambda_1 > \cdots > \lambda_N$ de paramètres de pénalité, avec $N \geq 1$. Soit s_k le nombre de composantes non nulles dans $\hat{\boldsymbol{\beta}}(\lambda_k)$, la solution obtenue pour le problème $\mathcal{P}(\lambda_k)$, et $S = \sum_{k=1}^N s_k$. La complexité globale de notre approche, sur l'ensemble des N valeurs consécutives, $(\lambda_k)_{k\in[N]}$ est $(2np + 7n + 11p + 12)N + 2nS + 4p(n + 1) + 2n$, ce qui est en général négligeable par rapport à la complexité des algorithmes de résolution du lasso. Cette complexité est de plus réduite si la matrice $\mathbf{X}$ est creuse. Enfin, au vu de (2.12), notre critère peut être calculé pour chaque variable indépendamment, sans avoir à charger la matrice $\mathbf{X}$ dans sa totalité, et notre approche est donc également facilement parallélisable.

Dans [VV4], nous évaluons SaFE sur des données réelles et des données simulées, notamment pour illustrer les problèmes de mémoire. Un premier point est que SaFe est particulièrement efficace à l'élimination de covariables pour les valeurs élevées du paramètre de pénalité λ. Une des applications pour lesquelles SaFE a été initialement développée consiste en l'analyse de grands corpus de documents et utilise des matrices d'occurrence de mots dans ces documents. Dans ce contexte, on est amené à chercher des solutions du lasso extrêmement creuses, même si cela signifie devoir travailler avec des valeurs du paramètre λ plus élevées que celles dictées par des critères liés au pouvoir prédictif par exemple. Nos résultats empiriques suggèrent que pour de telles valeurs du paramètre λ, SaFE permet une diminution importante du nombre de covariables, typiquement par un ordre de grandeur ou plus. Plus généralement, nos résultats empiriques suggèrent deux intérêts pratiques principaux de notre approche. Pour des matrices de design de taille modérée à grande, les temps nécessaires à la résolution numérique du lasso sont réduits lorsqu'on le combine à SaFE (ce qui est particulièrement intéressant lorsque plusieurs centaines de lasso doivent être résolus comme par exemple pour estimer la structure des modèles graphiques ; voir le chapitre 5). D'autre part, et peut-être surtout, SaFE étend la portée des algorithmes classiques de résolution du lasso en leur permettant de traiter des données de dimension tellement élevée qu'ils se heurtent sans SaFE à des problèmes de mémoire.

Des extensions de SaFE au cas du lasso avec intercept non pénalisé et de l'elastic net [Zou and Hastie, 2005] sont présentées dans [VV4], tout comme les extensions aux versions pénalisées par la norme L_1 de la régression logistique et des *support vector machines*. Dans ces deux derniers cas, l'analogue du problème d'optimisation (2.11) n'admet toutefois pas de forme analytique et doit être résolu numériquement.

L'approche décrite dans [VV4] était la première méthode de présélection à bénéficier de la propriété « safe ». Plusieurs travaux ont depuis étendu notre approche [Xiang et al., 2011, Xiang and Ramadge, 2012, Dai and Pelckmans, 2012, Wang et al., 2013, Xiang et al., 2014, Fercoq et al., 2015, Ndiaye et al., 2015]. En particulier, un champ de recherche s'intéresse

à l'incorporation de l'étape d'élimination des covariables au sein même de l'algorithme itératif de résolution du lasso. Les critères qui en résultent sont de type « dynamic safe rules » [Bonnefoy et al., 2014]. Une autre approche, similaire en principe à SaFE mais ne bénéficiant pas directement de la propriété safe, a été proposée à la suite de nos travaux par [Tibshirani et al., 2012]. Elle a été par la suite incorporée au package `glmnet` [Friedman et al., 2010], ce qui a très largement réduit les problèmes de mémoire de ce package, notamment dans le cas du modèle logistique.

Chapitre 3

Fused lasso généralisé : théorie asymptotique et robustesse à une mauvaise spécification du graphe

Dans la publication [VV11], nous étudions le fused lasso generalisé défini en (1.5). Pour rappel, la pénalité utilisée dans cette approche est double. La norme L_1 du vecteur des paramètres intervient afin d'encourager la sparsité des solutions. D'autre part, le terme de pénalité inclut également toutes les différences $|\beta_j - \beta_\ell|$ pour $j \sim \ell$, c'est-à-dire pour toute paire de covariables connectées dans un graphe donné a priori. Un graphe $G = (V, E)$ consiste en un ensemble de noeuds $V = \{1, \ldots, p\}$, qui correspond dans notre cas aux indices des composantes du vecteur $\boldsymbol{\beta}^*$ (et donc à l'ensemble des covariables de la matrice $\mathbf{X}$) et un ensemble d'arêtes E qui correspond aux paires d'indices (j, ℓ), $j > \ell$, des composantes connectées dans le graphe. Ce graphe décrit la structure attendue dans le vecteur des paramètres théoriques $\boldsymbol{\beta}^*$. Du point de vue de l'inférence, il fournit donc une information a priori. En tant que tel, le graphe peut être plus ou moins bien adapté aux données à analyser. Par exemple, le clustered lasso [She, 2010] correspond au fused lasso généralisé utilisant comme graphe la clique à p noeuds, c'est-à-dire le graphe dont l'ensemble E est l'ensemble des $p(p-1)/2$ arêtes possibles parmi les p noeuds. Le clustered lasso a été initialement proposé lorsque seule l'existence d'une structure en réseau est supposée, mais qu'aucune information n'est disponible sur la structure précise de ce réseau. Son terme de pénalité reposant sur toutes les différences, le clustered lasso pénalise généralement des différences correspondant à des composantes de valeurs distinctes dans $\boldsymbol{\beta}^*$. D'autre part, dans le cas où une information est disponible a priori sur la structure de $\boldsymbol{\beta}^*$, à partir d'une connaissance d'experts par exemple, cette information est rarement parfaite. Le graphe utilisé dans la pénalité, décrivant cette structure « pressentie », contient donc le plus souvent lui aussi des arêtes entre composantes de valeurs différentes, et en omet d'autres entre composantes de valeurs identiques. Ainsi, que l'on utilise la clique ou un graphe déterminé par un expert, la question de la robustesse du fused lasso généralisé se pose quant à une mauvaise spécification de ce graphe.

Dans [VV11], nous nous plaçons dans le cadre asymptotique en n, avec p fixe. Nous y établissons une propriété oraculaire asymptotique pour la version adaptative du fused lasso généralisé. Ce résultat établit notamment que deux composantes égales dans $\boldsymbol{\beta}^*$ seront estimées par une valeur commune avec probabilité qui tend vers 1 lorsque $n \to \infty$, si

elles appartiennent à la même composante connexe d'un sous-graphe de G, qui dépend de la structure de $\boldsymbol{\beta}^*$. En particulier, notre résultat établit que la version adaptative du clustered lasso (qui utilise la clique) est optimale asymptotiquement, lorsque p est supposé fixe. Nous associons à nos résultats théoriques une étude de simulation portant sur la robustesse du fused lasso généralisé face à une mauvaise spécification du graphe sur des échantillons de taille finie qui viennent tempérer les résultats asymptotiques en faveur de la stratégie utilisant la clique notamment. L'ensemble de ces résultats est résumé dans les paragraphes suivants. Ils complètent les résultats obtenus par [Sharpnack et al., 2012] et [Qian and Jia, 2016] sous le modèle de suite gaussienne tronquée, où $\mathbf{X} = \mathbf{I}_n$ et donc $p = n$ n'est pas fixe.

3.1 Résultats asymptotiques pour le fused lasso généralisé adaptatif

Pour simplifier l'exposé, nous nous plaçons une nouvelle fois sous le modèle linéaire (1.1). Les résultats présentés ici sont des versions simplifiées de certains des résultats de [VV11], qui sont eux établis dans le cadre des modèles linéaires généralisés.

Etant donné un graphe $G = (V, E)$ décrivant un a priori sur la structure du vecteur $\boldsymbol{\beta}^*$, nous nous intéressons à la version adaptative du fused lasso généralisé, qui reprend les idées du lasso adaptatif de [Zou, 2006] (voir l'annexe A). Comme nous nous plaçons dans le cadre asymptotique en n avec p fixe, nous utilisons des poids adaptatifs reposant sur l'estimateur des MCO $\widetilde{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}^*$. Pour un réel $\gamma > 0$ donné (par exemple $\gamma = 1$), on pose $w_{1,j} = |\widetilde{\beta}_j|^{-\gamma}$ et $w_{2,j,\ell} = |\widetilde{\beta}_j - \widetilde{\beta}_\ell|^{-\gamma}$ pour tout $(j, \ell) \in [p]^2$. Le fused lasso généralisé adaptatif se définit alors comme une solution du problème d'optimisation suivant, pour deux paramètres de régularisation λ_1, λ_2 positifs :

$$\text{minimiser}_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda_1 \sum_{j \in [p]} w_{1,j} |\beta_j| + \lambda_2 \sum_{(j,\ell) \in E} w_{2,j,\ell} |\beta_j - \beta_\ell|. \tag{3.1}$$

Ce critère est une variante du critère (1.5), où l'on utilise des versions pondérées des termes de pénalité. Les poids utilisés sont d'autant plus grands que les quantités $\tilde{\beta}_j$ et $\tilde{\beta}_j - \tilde{\beta}_\ell$ sont proches de 0. Plus précisément, avec probabilité tendant vers un, les poids associés aux quantités $|\beta_j|$ et $|\beta_j - \beta_\ell|$ tendent vers l'infini si $|\beta_j^*|$ et $|\beta_j^* - \beta_\ell^*|$ sont nulles, sous les hypothèses énoncées ci-dessous.

Dans ce chapitre, nous travaillerons sous les hypothèses suivantes, qui sont classiques pour l'étude asymptotique des estimateurs sous le modèle linéaire.

AGF1 Les variables ε_i, pour $i \in [n]$, sont i.i.d., d'espérance nulle et de variance $\sigma^2 > 0$.

AGF2 $(\mathbf{X}^T\mathbf{X})/n$ converge vers une matrice $\mathbf{C}$ définie positive lorsque $n \to \infty$.

Avant de présenter nos résultats théoriques, il nous faut introduire quelques notations. Comme précédemment, $J^* = \{j \in [p] : \beta_j^* \neq 0\}$ désigne le support du vecteur $\boldsymbol{\beta}^*$ et $p_0 = |J^*|$ son cardinal. On considère par ailleurs l'ensemble suivant de paires d'indices,

$$\mathcal{B} = \{(j, \ell) \in E,\ \beta_j^* \neq 0 \text{ et } \beta_j^* = \beta_\ell^*\} \subset J^* \times J^*.$$

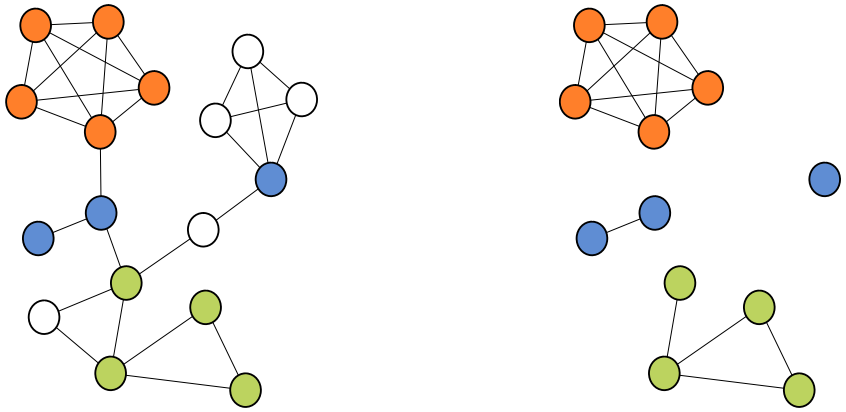


FIGURE 3.1 – *(A gauche)* Un exemple de graphe $G = (V, E)$ où la couleur des $p = 17$ noeuds indique la valeur du coefficient β_j^* correspondant. Les noeuds blancs correspondent à des composantes nulles, et les noeuds de même couleur à des composantes partageant la même valeur. *(A droite)* Le graphe $G_{\mathcal{B}} = (J^*, \mathcal{B})$ correspondant, où quatre composantes connexes apparaissent, $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$. Dans cet exemple, on a donc $s_0 = 4$, alors que $p_0 = 12$ et $d_0 = 3$. En particulier, $d_0 < s_0$ puisque le noeud bleu à droite n'est pas dans la même composante connexe que les deux noeuds bleus à gauche.

Après avoir observé que $J^* \subseteq V$ et $\mathcal{B} \subseteq E$, on définit $G_{\mathcal{B}}$ le sous-graphe de G tel que $G_{\mathcal{B}} = (J^*, \mathcal{B})$. Un exemple est donné en Figure 3.1. Ce graphe n'est bien sûr pas connu en pratique puisque J^* et $\mathcal{B}$ dépendent de $\boldsymbol{\beta}^*$, inconnu. Il joue cependant un rôle central sur les propriétés théoriques des estimateurs du fused lasso généralisé. En particulier, soit s_0 le nombre de composantes connexes de $G_{\mathcal{B}}$. La quantité s_0 peut être vue comme la complexité de $\boldsymbol{\beta}^*$ « portée » par G. On a clairement $d_0 \leq s_0 \leq p_0$, où d_0 est le nombre de valeurs distinctes non-nulles parmi les composantes de $\boldsymbol{\beta}^*$. On peut remarquer que $s_0 = p_0$ si et seulement si $(\beta_j^* = \beta_\ell^* \neq 0 \Rightarrow (j, \ell) \notin E)$. D'autre part, on a $s_0 = d_0$ si et seulement si, pour tout (j, ℓ) tel que $\beta_j^* = \beta_\ell^* \neq 0$, j et ℓ appartiennent à la même composante connexe de $G_{\mathcal{B}}$.

Pour tout $s \in [s_0]$, soit $\mathcal{A}_s \subset [p]$ l'ensemble des noeuds de la s-ème composante connexe de $G_{\mathcal{B}}$; en particulier $J^* = \bigcup_{s=1}^{s_0} \mathcal{A}_s$, et $\{\mathcal{A}_1, \ldots, \mathcal{A}_{s_0}\}$ est une partition de J^*. Notons par ailleurs $j_s = \min\{\mathcal{A}_s\}$ pour $s \in [s_0]$, et $\mathcal{A} = \{j_1, \ldots, j_{s_0}\}$. Après avoir rappelé que pour tout $s \in [s_0]$ et pour tout $j \in \mathcal{A}_s$, $\beta_j^* = \beta_{j_s}^*$, on définit $\boldsymbol{\beta}_{\mathcal{A}}^* = (\beta_{j_1}^*, \ldots, \beta_{j_{s_0}}^*)^T$ et $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{ad} = (\hat{\beta}_{j_1}^{ad}, \ldots, \hat{\beta}_{j_{s_0}}^{ad})^T$.

Soit alors $\tilde{\mathbf{X}}_{\mathcal{A}}$ la matrice de taille $n \times s_0$, dont la s-ème colonne est $\tilde{X}_s = \sum_{j \in \mathcal{A}_s} X_j$: la s-ème colonne de la matrice $\tilde{\mathbf{X}}_{\mathcal{A}}$ est donc la somme des colonnes de la matrice $\mathbf{X}$ correspondant aux indices de la s-ème composante connexe $\mathcal{A}_s$ (qui correspondent donc à des composantes de $\boldsymbol{\beta}^*$ égales entre elles, et non nulles). On introduit $\tilde{\mathbf{C}}_{\mathcal{A}}$ la matrice définie positive de taille $s_0 \times s_0$ définie comme la limite de $(\tilde{\mathbf{X}}_{\mathcal{A}}^T \tilde{\mathbf{X}}_{\mathcal{A}})/n$ lorsque $n \to \infty$. Finalement, soit

$\hat{J}_n = \{1 \leq j \leq p, \widehat{\beta}_j^{ad} \neq 0\}$ et, pour tout $s \in [s_0]$, $\mathcal{A}_{n,s} = \{\ell \in [p] : \hat{\beta}_\ell^{ad} = \hat{\beta}_{j_s}^{ad}\}$ (si bien que $j_s \in \mathcal{A}_{n,s}$). On peut maintenant présenter le résultat du théorème principal de [VV11], dans le cas du modèle linéaire.

Théorème 3.1.1 *Si $\lambda_m/\sqrt{n} \to 0$ et $\lambda_m n^{(\gamma-1)/2} \to \infty$, pour $m = 1, 2$, alors sous les hypothèses* **AGF1-2***, l'estimateur fused lasso généralisé adaptatif satisfait les propriétés suivantes :*

1. *Consistance en sélection de variables : lorsque $n \to +\infty$, on a $\mathbb{P}\big[\hat{J}_n = J^*\big] \to 1$ et, pour tout $s \in [s_0]$, $\mathbb{P}\left[\mathcal{A}_{n,s} = \mathcal{A}_s\right] \to 1$.*
2. *Normalité asymptotique : $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{ad} - \boldsymbol{\beta}_{\mathcal{A}}^*) \longrightarrow_d \mathcal{N}(\mathbf{0}_{s_0}, \sigma^2 \tilde{\mathbf{C}}_{\mathcal{A}}^{-1})$.*

Ce résultat établit une propriété oraculaire asymptotique du fused lasso généralisé. D'une part, le support J^* et chacune des composantes connexes $\mathcal{A}_s$ du graphe $G_{\mathcal{B}}$ sont identifiés avec une probabilité qui tend vers 1 lorsque $n \to \infty$. D'autre part, l'estimateur $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{ad}$ a la même loi limite que l'estimateur « oraculaire », c'est-à-dire celui des MCO construit à partir de la matrice de design $\tilde{\mathbf{X}}_{\mathcal{A}}$.

3.2 Interprétation et impact du graphe sur les performances

Le Théorème 3.1.1 nous permet également d'étudier l'impact du graphe utilisé dans la pénalité sur les propriétés asymptotiques de l'estimation, dans le cas où p est supposé fixe. En particulier, dès lors que $s_0 = d_0$, l'estimateur $\hat{\boldsymbol{\beta}}^{ad}$ a la même distribution asymptotique que l'estimateur oraculaire que l'on obtiendrait si l'on connaissait la vraie structure dans $\boldsymbol{\beta}^*$. C'est notamment le cas lorsqu'on utilise la clique. De plus, ajouter des arêtes entre des composantes de $\boldsymbol{\beta}^*$ de valeurs différentes ne modifie pas l'ensemble $\mathcal{B}$, et donc pas non plus la quantité s_0, alors qu'ajouter des arêtes entre des composantes de $\boldsymbol{\beta}^*$ de valeur identique fait croître l'ensemble $\mathcal{B}$ et peut faire diminuer s_0, et donc améliorer les performances asymptotiques du fused lasso généralisé adaptatif (en matière d'erreur de prédiction par exemple). A contrario, ôter des arêtes d'un graphe donné ne peut que faire croître la quantité s_0 (ou la laisser inchangée) : en particulier, éliminer des arêtes entre des composantes de $\boldsymbol{\beta}^*$ de même valeur fait décroître l'ensemble $\mathcal{B}$ et peut augmenter la quantité s_0, et donc dégrader les performances asymptotiques du fused lasso généralisé adaptatif. A l'extrême, le cas d'un graphe dont les arêtes ne connectent que des composantes distinctes de $\boldsymbol{\beta}^*$ correspond à $s_0 = p_0$ et revient donc au lasso adaptatif (qui correspond au fused lasso généralisé avec un graphe vide).

Bien sûr, ces résultats étant obtenus dans le cadre asymptotique avec p fixe, ils ne décrivent pas la réalité sur un échantillon de taille finie, ou face à des données de grande dimension. Dans le cas particulier du modèle de suite gaussienne tronquée, où $\mathbf{X} = \mathbf{I}_n$ et donc $n = p$, [Sharpnack et al., 2012] étudient des conditions portant sur le graphe sous lesquelles le fused lasso généralisé permet l'identification de la partition $\{\mathcal{A}_1, \ldots, \mathcal{A}_{s_0}\}$ avec probabilité qui tend vers 1. Même si l'extension des résultats de [Sharpnack et al., 2012] à des modèles plus généraux n'est pas triviale, ils suggèrent que le graphe utilisé doit être en

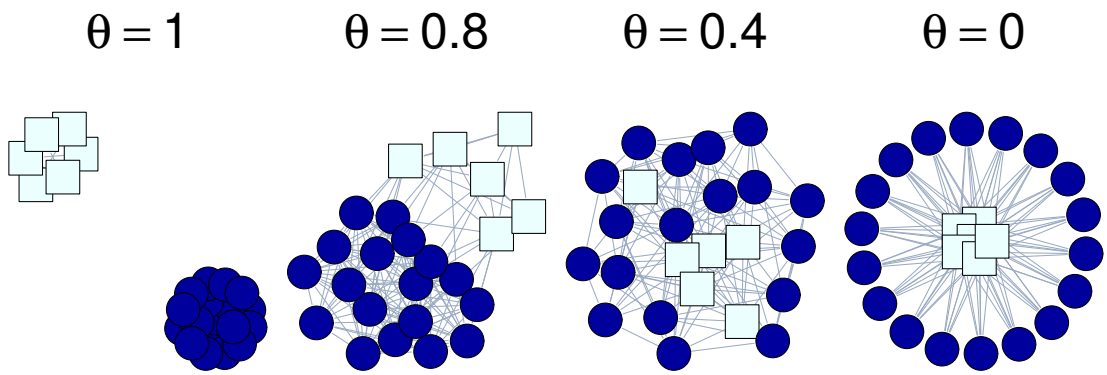


FIGURE 3.2 – Description de la génération des graphes utilisés dans la pénalité fused lasso généralisée en fonction du paramètre θ.

bonne adéquation avec la véritable structure du vecteur de paramètres théoriques $\boldsymbol{\beta}^*$ pour assurer cette identification.

Nous avons effectué une étude de simulation approfondie pour comparer les performances du fused lasso généralisé et d'autres approches pénalisées. Notre objectif principal était d'étudier l'apport de la prise en compte de la structure attendue de $\boldsymbol{\beta}^*$ décrite par le graphe, en fonction notamment de son adéquation avec la véritable structure de $\boldsymbol{\beta}^*$. Pour ce faire, nous avons calculé les estimateurs du fused lasso généralisé en faisant varier le graphe dans la pénalité : nous considérons la clique, le graphe vide (auquel cas le fused lasso généralisé revient à un simple lasso), et quatre graphes générés aléatoirement et dépendant d'un paramètre mesurant l'adéquation à la véritable structure du vecteur $\boldsymbol{\beta}^*$. Plus précisément, pour un vecteur $\boldsymbol{\beta}^* \in \mathbb{R}^p$ donné, dont $p/2$ composantes sont nulles et les $p/2$ restantes égales à un réel $\beta^* > 0$ donné, nous avons généré des graphes tels que les paires d'indices correspondant à des composantes de $\boldsymbol{\beta}^*$ de même valeur sont connectées avec probabilité θ, et les paires correspondant à des composantes de valeurs distinctes avec probabilité $1 - \theta$. Une illustration est donnée en Figure 3.2. Lorsque $\theta = 1$, le graphe est en parfaite adéquation avec la structure de $\boldsymbol{\beta}^*$ puisque les composantes égales à β^* forment une clique, les composantes nulles en forment une autre, et ces deux cliques ne sont pas connectées entre elles.

Concernant la calibration des paramètres de régularisation λ_1 et λ_2, nous avons opté pour des critères de type 2stepBIC (voir le paragraphe A.5 de l'annexe A). Ils sont en effet adaptés lorsque la question d'intérêt porte sur la sélection des variables, et plus généralement la structure de $\boldsymbol{\beta}^*$, et lorsque p est petit devant n (qui est le cas dans nos simulations). Nous comparons les versions 0-relaxées des différentes approches (voir le paragraphe A.4 de l'annexe A). Pour le lasso, il s'agit donc de la version OLS-Hybrid, et pour le fused lasso généralisé de son extension naturelle.

Dans le cas d'un graphe bien adapté à la véritable structure de $\boldsymbol{\beta}^*$ ($\theta \geq 0.8$), le fused

lasso généralisé 0-relaxé, dans sa version standard ou adaptative, surpasse nettement le lasso 0-relaxé en matière de sélection du support et de pouvoir prédictif, ce qui illustre l'effet « coopératif » engendré par la pénalité de type fused : en particulier, le fused lasso détecte plus précisément le support de $\boldsymbol{\beta}^*$ pour des signaux faibles grâce aux arêtes qui connectent dans le graphe les composantes de $\boldsymbol{\beta}^*$ de même valeur. D'autre part, lorsque l'adéquation entre le graphe et la structure de $\boldsymbol{\beta}^*$ diminue, les performances du fused lasso généralisé diminuent également. Elles peuvent même être moindres que celles du lasso, notamment pour l'identification du support, mais nos résultats suggèrent que le fused lasso généralisé fait toujours au moins aussi bien que le lasso d'un point de vue du pouvoir prédictif. Nos résultats suggèrent également que la version adaptative du fused lasso généralisé est plus robuste à une mauvaise spécification du graphe et que la 0-relaxation améliore elle aussi la robustesse de l'approche. D'autre part, concernant la stratégie utilisant la clique, nous observons des performances proches de celles obtenues pour des graphes faiblement adaptés au vecteur $\boldsymbol{\beta}^*$, correspondant à des valeurs de $\theta \in [0, 0.4]$ sur les configurations considérées dans nos simulations. Sur ces configurations, le fused lasso généralisé utilisé avec la clique montre des performances similaires à celles du lasso, quant à la sélection du support ou le pouvoir prédictif, avec l'avantage bien sûr d'identifier certaines paires de composantes partageant la même valeur. Ces résultats complètent ainsi ceux du Théorème 3.1.1 ci-dessus : même si la clique est optimale dans un cadre asymptotique en n (et où p est supposé fixe), elle est généralement sous-optimale sur des échantillons de taille finie. Ils confirment ainsi l'intuition selon laquelle les performances du fused lasso généralisé sont accrues si le graphe fourni par les experts est bien adapté à la vraie structure de $\boldsymbol{\beta}^*$.

Dans [VV11], nous évaluons de plus les propriétés du fused lasso généralisé lorsque les composantes non-nulles de $\boldsymbol{\beta}^*$ ne sont pas nécessairement strictement égales, en considérant le cas où chacune de ces composantes est générée aléatoirement selon $\beta^* + \nu$ où $\nu \sim \mathcal{N}(0, \sigma_\nu^2)$, avec $\sigma_\nu^2 \in \{0, 0.2, 0.5\}$. Le vecteur $\boldsymbol{\beta}^*$ est alors composé d'un groupe de composantes nulles, et d'un groupe de composantes non-nulles (et de valeurs plus ou moins proches les unes des autres). Nous avons comparé les performances du fused lasso généralisé à celles du group lasso [Yuan and Lin, 2006], qui constitue une option naturelle dans cette situation, mais qui nécessite la connaissance a priori des deux groupes. Nos résultats indiquent qu'en supposant la connaissance a priori des groupes (et en utilisant donc un graphe parfaitement adapté), le fused lasso généralisé surpasse le plus souvent le group lasso, en matière de détection du support et de pouvoir discriminant. D'autre part, même si les groupes ne sont pas connus exactement lors de l'application du fused lasso généralisé (et donc le graphe utilisé n'est pas parfaitement adapté), le fused lasso généralisé fait souvent aussi bien, voire mieux, que le group lasso (qui lui repose sur la connaissance exacte des groupes). Ces résultats empiriques suggèrent que le fused lasso généralisé est une approche à considérer lorsqu'un graphe est disponible et décrit les similarités attendues entre les composantes de $\boldsymbol{\beta}^*$ plutôt que des égalités strictes.

Deuxième partie

Approches pénalisées pour données stratifiées

Chapitre 4

Modèles de régression sur données stratifiées

4.1 Introduction

Comme nous l'avons montré dans le chapitre introductif de ce document, il est fréquent, en épidémiologie notamment, que les observations proviennent de différents sous-groupes d'une population. Ces sous-groupes, ou *strates*, sont généralement définis à travers les niveaux d'une variable catégorielle Z telle que le sexe de l'individu, le dosage ou le type de traitement, la zone géographique, etc. ou des combinaisons de ces variables.

Dans ce chapitre, nous nous intéressons à l'étude de l'association entre une variable d'intérêt Y et un vecteur de covariables $\mathbf{x}$. L'objectif principal est alors de décrire comment la variable Z influe sur l'association entre les covariables $\mathbf{x}$ et la variable d'intérêt Y. Pour simplifier l'exposé, les méthodes et résultats seront une nouvelle fois présentés dans le cas de la régression linéaire homoscédastique sur designs déterministes.

Reprenons pour commencer les notations introduites dans le chapitre introductif. Nous supposons disposer d'un n-échantillon, $n \geq 1$, tel que l'observation $i \in [n]$ correspond au triplet $(Y_i, \mathbf{x}_i, Z_i)$ où $Y_i \in \mathbb{R}$ est la variable d'intérêt, $\mathbf{x}_i \in \mathbb{R}^p$ le vecteur des covariables, et $Z_i \in [K]$ la variable catégorielle décrivant la strate d'appartenance de l'observation i. Soit $n_k = \sum_{i \in [n]} \mathbb{I}(Z_i = k)$, le nombre d'observations de la strate k, si bien que $n = \sum_{k \in [K]} n_k$. Pour tout $k \in [K]$, on définit $\mathbf{Y}^{(k)} = (y_1^{(k)}, \ldots, y_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$ le vecteur de variables réponse et $\mathbf{X}^{(k)} = ({\mathbf{x}_1^{(k)}}^T, \ldots, {\mathbf{x}_{n_k}^{(k)}}^T)^T \in \mathbb{R}^{n_k \times p}$ la matrice de design correspondant aux observations de la strate k, c'est-à-dire aux observations $i \in [n]$ telles que $Z_i = k$. On définit par ailleurs $\boldsymbol{\varepsilon}^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$ le vecteur des résidus dans cette strate, dont on supposera qu'il vérifie $\mathbb{E}\boldsymbol{\varepsilon}^{(k)} = \mathbf{0}_{n_k}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}^{(k)}) = \sigma^2 \mathbf{I}_{n_k}$. On considérera alors que les vecteurs $\mathbf{Y}^{(k)}$ sont liés aux matrices de design $\mathbf{X}^{(k)}$ à travers les modèles de régression linéaire suivants, décrivant chacun l'association entre Y et $\mathbf{x}$ sur chacune des K strates :

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)} \boldsymbol{\beta}_k^* + \boldsymbol{\varepsilon}^{(k)} \quad \text{pour tout } k \in [K], \tag{4.1}$$

où les vecteurs de paramètres $\boldsymbol{\beta}_k^*$ sont fixes mais inconnus.

Une stratégie classique consiste à estimer les K vecteurs $\boldsymbol{\beta}_k^*$ de manière indépendante. On peut par exemple résoudre un lasso sur chaque strate, pour sélectionner les covariables

associées à la variable d'intérêt sur chaque strate, et estimer leurs effets. Cette stratégie sera désignée par IndepLasso ci-dessous. Elle revient à définir $\hat{\boldsymbol{\beta}}_k$, pour tout k, comme solution minimisant le critère suivant

$$\frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2} + \lambda_k\|\boldsymbol{\beta}_k\|_1,$$

pour des paramètres de régularisation $\lambda_k \geq 0$ donnés, $k \in [K]$. Notons pour la suite que les solutions $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K)$ retournées par IndepLasso s'obtiennent de manière équivalente comme solution minimisant le critère

$$\sum_{k\in[K]} \left\{ \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2} + \lambda_k\|\boldsymbol{\beta}_k\|_1 \right\}. \tag{4.2}$$

Une seconde stratégie consisterait à négliger l'information relative aux strates, et travailler implicitement sous l'hypothèse $\boldsymbol{\beta}_1^* = \ldots = \boldsymbol{\beta}_K^*$, et donc sous le modèle

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^{(k)} \quad \text{pour tout } k \in [K]. \tag{4.3}$$

Par exemple, ce que nous désignerons par PoolLasso consiste à minimiser en $\boldsymbol{\beta} \in \mathbb{R}^p$ le critère suivant, pour un paramètre de régularisation $\lambda \geq 0$ donné,

$$\sum_{k\in[K]} \left\{ \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}\|_2^2}{2} \right\} + \lambda\|\boldsymbol{\beta}\|_1. \tag{4.4}$$

Le point commun de ces deux approches est qu'elles ne permettent pas de s'adapter au niveau réel, mais inconnu, d'homogénéité entre les vecteurs $\boldsymbol{\beta}_k^*$. IndepLasso ne tire ainsi aucunement profit de l'homogénéité éventuelle entre les vecteurs $\boldsymbol{\beta}_k^*$, $k \in [K]$, et renvoie donc typiquement des estimations de variance inutilement grande. A contrario, PoolLasso masque toute hétérogénéité éventuelle entre les vecteurs $\boldsymbol{\beta}_k^*$, $k \in [K]$, et renvoie donc des estimations typiquement biaisées. Outre leurs défauts respectifs quant à la qualité de l'estimation, ces deux stratégies ne permettent pas d'étudier le rôle de la variable Z sur l'association entre Y et $\mathbf{x}$. Avec IndepLasso en particulier, on ne peut pas interpréter les différences observées entre $\hat{\beta}_{k_1,j}$ et $\hat{\beta}_{k_2,j}$ pour deux strates $k_1 \neq k_2$ et $j \in [p]$ fixé, puisque ces valeurs sont différentes par construction.

Une autre stratégie classique en épidémiologie a été brièvement présentée dans le chapitre introductif. Elle consiste à sélectionner une strate de référence ℓ, *a priori*, puis à décomposer les paramètres du modèle (4.1) selon l'équation $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_\ell^* + \boldsymbol{\delta}_k^*$, pour tout $k \in [K]$, avec $\boldsymbol{\delta}_\ell^* = \mathbf{0}_p$ [Gertheiss and Tutz, 2012]. On peut une nouvelle fois appliquer le lasso pour estimer et sélectionner les paramètres sous cette nouvelle paramétrisation, c'est-à-dire pour déterminer quelles composantes du vecteur $\boldsymbol{\beta}_\ell^*$ d'une part, et des vecteurs $\boldsymbol{\delta}_k^*$ d'autre part, sont nulles. Plus précisément, les estimateurs $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K)$ sont obtenus à partir des solutions qui minimisent le critère suivant, pour des valeurs positives données des paramètres λ_1 et $\lambda_{2,k}$:

$$\frac{1}{2}\Big\{\|\mathbf{Y}^{(\ell)} - \mathbf{X}^{(\ell)}\boldsymbol{\beta}_\ell\|_2^2 + \sum_{k\neq\ell}\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\beta}_\ell + \boldsymbol{\delta}_k)\|_2^2\Big\} + \lambda_1\|\boldsymbol{\beta}_\ell\|_1 + \sum_{k\neq\ell}\lambda_{2,k}\|\boldsymbol{\delta}_k\|_1. \tag{4.5}$$

Par la suite, on désigne cette approche par RefLasso. Elle tire en partie profit de l'homogénéité éventuelle entre les vecteurs $\boldsymbol{\beta}_k^*$. On peut espérer qu'elle permette d'étudier le rôle de la variable Z sur l'association entre Y et $\mathbf{x}$, mais seulement en partie. En effet, seules les différences entre les effets des covariables sur la strate de référence de ℓ et les autres strates sont pénalisées, et les différences des effets des covariables entre deux strates $k_1 \neq \ell$ et $k_2 \neq \ell$ ne le sont pas. On ne peut donc pas interpréter les différences éventuelles entre $\hat{\beta}_{k_1,j}$ et $\hat{\beta}_{k_2,j}$ en termes d'effet de Z sur l'association entre Y et la j-ème covariable s'ils sont tous deux différents de $\hat{\beta}_{\ell,j}$.

Du point de vue de la qualité de l'estimation, un deuxième défaut de RefLasso vient du fait que le nombre de paramètres non nuls dans le modèle reparamétré suite au choix ℓ de la strate de référence vaut $\|\boldsymbol{\beta}_\ell^*\|_0 + \sum_{k\neq\ell} \|\boldsymbol{\delta}_k^*\|_0$. Cette dimension est minimale si la strate de référence ℓ est telle que $\beta_{\ell,j}^*$ est un des modes de l'ensemble des valeurs $(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$, et ce pour tout $j \in [p]$. Une strate $\ell \in [K]$ telle que $\beta_{\ell,j}^* \in \text{mode}(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$ pour tout $j \in [p]$ n'existe que rarement en pratique. Par contre, pour tout $j \in [p]$, il existe toujours (au moins) une strate $\ell_j^* \in [K]$ telle que $\beta_{\ell_j^*}^* \in \text{mode}(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$. Si une telle strate de référence « covariable-dépendante » ℓ_j^* était connue pour tout $j \in [p]$, alors une alternative à RefLasso consisterait à utiliser la paramétrisation $\beta_{k,j}^* = \beta_{\ell_j^*,j}^* + \tilde{\delta}_{k,j}^*$, pour tout $k \neq \ell_j^*$ avec $\tilde{\delta}_{k,j}^* = \beta_{k,j}^* - \beta_{\ell_j^*,j}^*$. La stratégie correspondante est oraculaire (au sens où elle nécessite l'intervention d'un oracle qui fournirait les strates covariable-dépendantes) et sera désignée par ORefLasso dans la suite. Evidemment en pratique, les strates ℓ_j^* ne sont généralement pas accessibles et il n'est donc pas possible d'appliquer ORefLasso. Nous reviendrons sur les performances relatives de RefLasso et ORefLasso, en matière de sélection de variables, dans le paragraphe 4.3.

Une famille de stratégies moins classiques en épidémiologie cherche à tirer profit de l'homogénéité éventuelle des $\boldsymbol{\beta}_k^*$, et plus précisément d'un certain type de *structure* attendu dans la matrice $\mathbf{B}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$. Ces stratégies sont issues de la littérature traitant de l'apprentissage multi-tâches [Evgeniou and Pontil, 2004, Argyriou et al., 2008], dont le problème de l'estimation simultanée des K modèles de régression (4.1) est un cas particulier. On peut citer les travaux de [Lounici et al., 2011] et [Negahban and Wainwright, 2011] qui étudient les propriétés d'estimateurs de deux versions du group-lasso (L_1/L_2 et L_1/L_∞) dans un cadre non-asymptotique, ou encore les travaux de [Maurer and Pontil, 2013] concernant une procédure reposant sur la norme nucléaire de la matrice des paramètres. Chacune de ces méthodes encourage ainsi un certain type de structure dans la matrice estimée $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K) \in \mathbb{R}^{p\times K}$. La norme nucléaire encourage cette matrice à être de rang faible alors que les deux approches de type group lasso encouragent une structure de sparsité au niveau des lignes de $\hat{\mathbf{B}}$: certaines lignes de $\hat{\mathbf{B}}$ sont « uniformément » nulles et les variables correspondantes ont un effet estimé nul sur l'ensemble des strates. La version L_1/L_∞ du group lasso encourage de plus les effets non nuls d'une variable à être égaux en valeur absolue sur les différentes strates : typiquement, les solutions sont telles qu'il existe $(k_1, k_2) \in [K]^2$, avec $k_1 \neq k_2$, et $j \in [p]$ avec $|\hat{\beta}_{k_1,j}| = |\hat{\beta}_{k_2,j}|$. Cette dernière propriété est particulièrement intéressante en vue de l'étude de l'effet de Z sur l'association entre Y et $\mathbf{x}$. En effet, puisque l'approche L_1/L_∞ encourage les solutions telles que $|\hat{\beta}_{k_1,j}| = |\hat{\beta}_{k_2,j}|$, si la solu-

tion finalement retournée est telle que $|\hat{\beta}_{k_1,j}| \neq |\hat{\beta}_{k_2,j}|$, alors cela suggère que $|\beta^*_{k_1,j}| \neq |\beta^*_{k_2,j}|$ et l'on peut donc interpréter ce résultat en termes d'effet de Z sur l'association entre Y et x_j. A contrario, il est impossible d'interpréter les différences $\hat{\beta}_{k_1,j} \neq \hat{\beta}_{k_2,j}$ pour j fixé en termes d'effet de Z sur l'association entre Y et x_j avec l'approche L_1/L_2 puisque chaque variable est sélectionnée de manière globale, et les effets estimés sur les différentes strates d'une variable globalement sélectionnée sont tous différents, par construction.

Puisque déterminer la façon dont Z modifie les effets des covariables revient à identifier, pour tout $j \in [p]$, les paires $(k_1, k_2) \in [K] \times [K]$ telles que $\beta^*_{k_1,j} = \beta^*_{k_2,j}$, je me suis particulièrement intéressé à des approches pénalisées encourageant les égalités du type $\hat{\beta}_{k_1,j} = \hat{\beta}_{k_2,j}$ à j fixé (contrairement à l'approche L_1/L_∞ qui encourage « seulement » les égalités en valeur absolue), et permettent ainsi d'interpréter les différences obtenues en termes d'effet de Z sur l'association entre Y et $\mathbf{x}$. Les paragraphes suivants décrivent l'utilisation du fused lasso généralisé dans ce contexte, puis une nouvelle approche, AutoRefLasso, qui peut être considérée comme une amélioration de RefLasso présentée ci-dessus.

4.2 Le fused lasso généralisé pour les données stratifiées

4.2.1 Principe général

Afin d'encourager les égalités du type $\hat{\beta}_{k,j} = \hat{\beta}_{\ell,j}$ à j fixé, il est relativement naturel de considérer la stratégie qui retourne les estimateurs $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K$ comme solutions minimisant le critère suivant :

$$\sum_k \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2} + \lambda_1 \sum_k \|\boldsymbol{\beta}_k\|_1 + \lambda_2 \sum_{\substack{(k,\ell)\in[K]^2 \\ k<\ell}} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_1. \tag{4.6}$$

Le terme $\sum_k \|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2/2$ mesure l'adéquation aux données. Le terme $\sum_k \|\boldsymbol{\beta}_k\|_1$ encourage les solutions $\hat{\boldsymbol{\beta}}_k$ à être creuses (sélection des variables au sein de chaque strate). Le terme $\sum_{k<\ell} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_1$ encourage finalement l'homogénéité des vecteurs solutions $\hat{\boldsymbol{\beta}}_k$, c'est-à-dire les solutions telles que $\hat{\beta}_{k,j} = \hat{\beta}_{\ell,j}$ pour $k \neq \ell$ et j fixé. La différence entre (4.6) et l'écriture (4.2) du critère d'IndepLasso réside dans le terme $\lambda_2 \sum_{k<\ell} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_1$. Alors qu'IndepLasso revient à résoudre les K problèmes lasso de manière indépendante, ce terme additionnel a pour effet de coupler les estimations des vecteurs $\boldsymbol{\beta}^*_k$, en les encourageant à être proches les unes des autres (du point de vue de la norme L_1).

Cette approche a été initialement proposée par [Gertheiss and Tutz, 2012] (voir aussi [Oelker et al., 2014]). Dans [VV11], nous montrons que le critère (4.6) est un cas particulier du critère minimisé par le fused lasso généralisé [Höfling et al., 2010], décrit au chapitre précédent. Soit $\mathscr{Y} = ({\mathbf{Y}^{(1)}}^T, \ldots, {\mathbf{Y}^{(K)}}^T) \in \mathbb{R}^n$ le vecteur renfermant les n observations de la variable réponse (sur l'ensemble des strates). Soit de plus $\boldsymbol{\mathcal{X}}_F$ la matrice diagonale par blocs de taille $(n \times Kp)$, dont le k-ème bloc est de dimension $n_k \times p$ et vaut $\mathbf{X}^{(k)}$ pour $k \in [K]$. Posons $\mathbf{b}^* = ({\boldsymbol{\beta}^*_1}^T, \ldots, {\boldsymbol{\beta}^*_K}^T)^T = (b^*_1, \ldots, b^*_{Kp}) \in \mathbb{R}^{Kp}$. Ici, les similarités attendues sont entre les composantes $\beta^*_{k_1,j}$ et $\beta^*_{k_2,j}$, pour $k_1 \neq k_2 \in [K]$ et $j \in [p]$, c'est-à-dire entre les composantes $j_1 \neq j_2 \in [Kp]$ du vecteur $\mathbf{b}^*$ telles que $j_1 \%\%p = j_2 \%\%p$, où $n_1 \%\% n_2$

désigne le reste de la division euclidienne de n_1 par n_2 pour tout couple d'entiers (n_1, n_2). Ainsi, en posant $E_C = \{(j_1, j_2) \in [Kp]^2 : j_1 \neq j_2, j_1 \%\% p = j_2 \%\% p\}$, on peut définir le graphe $\mathcal{G}_C = (V_C, E_C)$ à Kp sommets, représentés par l'ensemble V_C qui contient les Kp composantes du vecteur $\mathbf{b}$, et dont l'ensemble des arêtes est E_C. Ce graphe est composé de p cliques de taille K (voir l'illustration en Figure 4.1-a, page 44) : une clique par covariable, la j-ème clique, $j \in [p]$, reliant entre elles l'ensemble des composantes de $\mathbf{b}^*$ qui correspondent aux paramètres $\beta^*_{1,j}, \ldots, \beta^*_{K,j}$. Etant donné ce graphe, le critère en (4.6) s'écrit comme celui d'un fused lasso généralisé :

$$\frac{\|\mathscr{Y} - \boldsymbol{\mathcal{X}}_F \mathbf{b}\|_2^2}{2} + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{(j_1,j_2)\in E_C} |b_{j_1} - b_{j_2}|, \tag{4.7}$$

que l'on cherche à minimiser en $\mathbf{b} \in \mathbb{R}^{Kp}$. Compte tenu de la forme particulière du graphe sur lequel repose cette stratégie, nous la désignerons par CliqueFused dans la suite de ce document.

4.2.2 Optimalité asymptotique de la version adaptative

Dans le cadre asymptotique en n, supposons que Kp est fixe et que les tailles n_k de chacune des strates croissent vers l'infini à la même vitesse, c'est-à-dire

$$\forall k \in [K], \exists \rho_k \in (0,1) : n_k/n \to \rho_k \text{ lorsque } n \to \infty.$$

Supposons de plus que pour tout $k \in [K]$, la matrice $({\mathbf{X}^{(k)}}^T \mathbf{X}^{(k)})/n_k$ converge vers une matrice définie positive $\mathbf{C}^{(k)}$ lorsque $n_k \to \infty$. On suppose enfin que les variables $\varepsilon_i^{(k)}$, pour tout $i \in [n_k]$ et $k \in [K]$ sont i.i.d., d'espérance nulle et de variance $\sigma^2 > 0$. Soit alors $(\tilde{\beta}_{k,j})_{k\in[K],j\in[p]}$ les estimations obtenues par la méthode des moindres carrés ordinaires, appliquée indépendamment sur chaque strate. La version adaptative de CliqueFused revient à définir les estimateurs $\hat{\boldsymbol{\beta}}_1^{ad}, \ldots, \hat{\boldsymbol{\beta}}_K^{ad}$ comme solution minimisant le critère suivant :

$$\sum_k \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \sum_{k\in[K]} \sum_{j\in[p]} \frac{|\beta_{k,j}|}{|\tilde{\beta}_{k,j}|^\gamma} + \lambda_2 \sum_{k_1<k_2} \sum_{j\in[p]} \frac{|\beta_{k_1,j} - \beta_{k_2,j}|}{|\tilde{\beta}_{k_1,j} - \tilde{\beta}_{k_2,j}|^\gamma}, \tag{4.8}$$

où $\gamma > 0$ est fixé (on prend typiquement $\gamma = 1$).

En supposant que K et p sont fixes (par rapport à n), le théorème 3.1.1 présenté au chapitre 3 permet d'établir une propriété oraculaire asymptotique pour la version adaptative de CliqueFused. Ce résultat est analogue à ceux obtenus dans [Gertheiss and Tutz, 2012] sous le modèle de régression linéaire et [Oelker et al., 2014] sous les modèles linéaires généralisés (notre théorème présenté dans [VV11] couvre également les modèles linéaires généralisés). Pour tout $j \in [p]$, soit $K_j^* = \{k \in [K] : \beta^*_{k,j} \neq 0\}$ et $0 \leq d_j \leq K$ le nombre de valeurs distinctes non nulles parmi l'ensemble des paramètres $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$. Si $d_j > 0$, ce qui revient à dire que $K_j^* \neq \emptyset$, on note $\mathcal{K}_j^* = (\kappa_j^{(1)}, \ldots, \kappa_j^{(d_j)})$ la partition de l'ensemble K_j^* telle

que pour tout $(k_1, k_2) \in [K]^2, \beta^*_{k_1,j} = \beta^*_{k_2,j} \neq 0 \Leftrightarrow \exists d \in [d_j] : (k_1, k_2) \in \kappa_j^{(d)}$. Si $d_j > 0$, soit $k_j^{(m)} = \min\{\kappa_j^{(m)}\}$ pour tout $m \in [d_j]$ et

$$\mathbf{b}^*_j = (\beta^*_{k_j^{(1)},j}, \ldots, \beta^*_{k_j^{(d_j)},j})$$

l'ensemble des d_j valeurs distinctes non nulles parmi $(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$. Observons que la connaissance des partitions $\mathcal{K}^*_j$ et des vecteurs $\mathbf{b}^*_j$ pour tout $j \in [p]$ décrit complètement l'effet de Z sur l'association entre Y et $\mathbf{x}$. On note $\mathbf{b}^*_{\mathcal{A}} = (\mathbf{b}^*_j)_{j:d_j>0}$ la concaténation des vecteurs $\mathbf{b}^*_j$ pour $j \in [p]$ tels que $d_j > 0$.

Pour tout $j \in [p]$, soit $\hat{K}_j = \{k \in [p] : \hat{\beta}^{ad}_{k,j} \neq 0\}$, et pour tout j tel que $d_j > 0$, $\hat{\mathbf{b}}_j = (\hat{\beta}^{ad}_{k_j^{(1)},j}, \ldots, \hat{\beta}^{ad}_{k_j^{(d_j)},j})$. On note $\hat{\mathbf{b}}^{ad}_{\mathcal{A}} = (\hat{\mathbf{b}}_j)_{j:d_j>0}$ la concaténation des vecteurs $\hat{\mathbf{b}}_j$ pour $j \in [p]$ tels que $d_j > 0$.

Nous devons maintenant définir la matrice $\tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{A}}$, de taille $n \times d_0$, avec $d_0 = \sum_j d_j$, qui correspond à la matrice que l'on utiliserait naturellement dans ce contexte si un oracle nous donnait les partitions $\mathcal{K}^*_j$ (et donc les ensembles K^*_j). Etant donné cette information, on éliminerait les colonnes de la matrice $\boldsymbol{\mathcal{X}}_F$ correspondant aux observations de la variable j sur les strates appartenant à K^{*c}_j, et on sommerait les colonnes de $\boldsymbol{\mathcal{X}}_F$ correspondant, pour une variable j donnée, aux strates appartenant à $\kappa_j^{(m)}$, pour chaque $m \in [d_j]$. Plus formellement, pour tout $j \in [p]$ tel que $d_j > 0$ et tout $m \in [d_j]$, soit $\mathcal{A}_j^{(m)}$ l'ensemble des indices de colonnes de la matrice $\boldsymbol{\mathcal{X}}_F$ correspondant aux observations de la j-ème variable dans les strates appartenant à $\kappa_j^{(m)}$: $\mathcal{A}_j^{(m)} = \{j_1 \in [Kp] : \exists k \in \kappa_j^{(m)} : j_1 = (k-1)p + j\}$. Soit alors, pour tout $j \in [p]$ tel que $d_j > 0$, $\mathcal{A}_j = \{\mathcal{A}_j^{(1)}, \ldots, \mathcal{A}_j^{(d_j)}\}$. On définit maintenant $\tilde{\boldsymbol{\mathcal{X}}}^{(j)}$ la matrice de taille $n \times d_j$, dont la m-ème colonne est donnée par $\sum_{j_1 \in \mathcal{A}_j^{(m)}} \mathcal{X}_{F,j_1}$: cette m-ème colonne est bien la somme des colonnes de la matrice $\boldsymbol{\mathcal{X}}_F$ qui correspondent aux observations de la j-ème variable dans le sous-ensemble de strates $\kappa_j^{(m)}$. La matrice $\tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{A}}$ est obtenue en concaténant en colonne les matrices $\tilde{\boldsymbol{\mathcal{X}}}^{(j)}$ pour tout j tel que $d_j > 0$. Finalement, soit $\tilde{\mathbf{C}}_{\mathcal{A}}$ la matrice définie positive de taille (d_0, d_0) définie comme la limite de $(\tilde{\boldsymbol{\mathcal{X}}}^T_{\mathcal{A}} \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{A}})/n$ lorsque $n \to \infty$. On peut maintenant énoncer le résultat suivant.

Corollaire 4.2.1 *Si $\lambda_m/\sqrt{n} \to 0$ et $\lambda_m n^{(\gamma-1)/2} \to \infty$, pour $m = 1, 2$, alors l'estimateur CliqueFused adaptatif satisfait les propriétés suivantes :*

1. *Consistance en sélection de variables : lorsque $n \to +\infty$, on a*

$$\mathbb{P}\big(\cap_{j\in[p]} \{K^*_j = \hat{K}_j\}\big) \to 1.$$

2. *Consistance pour la détection des hétérogénéités : pour tout $j \in [p]$, on a, avec probabilité qui tend vers 1 lorsque $n \to +\infty$:*

$$\beta^*_{k_1,j} = \beta^*_{k_2,j} \Leftrightarrow \hat{\beta}^{ad}_{k_1,j} = \hat{\beta}^{ad}_{k_2,j}.$$

3. *Normalité asymptotique : $\sqrt{n}(\hat{\mathbf{b}}^{ad}_{\mathcal{A}} - \mathbf{b}^*_{\mathcal{A}}) \longrightarrow_d \mathcal{N}(\mathbf{0}_{s_0}, \sigma^2 \tilde{\mathbf{C}}^{-1}_{\mathcal{A}})$.*

Ce corollaire établit notamment que pour chaque covariable $j \in [p]$, l'ensemble $K_j^* = \{k \in [K] : \beta_{k,j}^* \neq 0\}$ et la partition $\mathcal{K}_j^*$ sont identifiés avec probabilité qui tend vers 1 lorsque $n \to \infty$. Il établit de plus que si $d_j > 0$, alors les estimateurs des effets $\beta_j^{(d)*}$ sur chaque sous-ensemble de strates $\kappa_j^{(d)} \subseteq [K]$, pour $d \in [d_j]$, ont la même loi limite que l'estimateur oraculaire qu'on obtiendrait en regroupant les observations issues de ces strates pour cette covariable, c'est-à-dire en travaillant avec la matrice $\tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{A}}$.

Ainsi, la version adaptative de CliqueFused permet, asymptotiquement et en supposant Kp fixe, de décrire précisément les hétérogénéités dans les vecteurs $\boldsymbol{\beta}_k^*$ et donc l'effet de la variable Z sur l'association entre Y et $\mathbf{x}$. Asymptotiquement, cette version adaptative conduit ainsi à l'estimation d'un nombre de paramètres minimal, compte tenu de ces hétérogénéités. Elle est optimale pour l'estimation d'un modèle de régression linéaire (voire linéaire généralisé) sur données stratifiées dans le cadre asymptotique en n, lorsque Kp est supposé fixe.

4.2.3 Extension aux modèles non linéaires à effets mixtes

Dans [VV7], nous étendons le fused lasso généralisé au cas des modèles non linéaires à effets mixtes, qui sont particulièrement utilisés en pharmacocinétique pour modéliser par exemple la quantité de médicament présente dans le sang en fonction du temps. Le fused lasso généralisé est notamment utile dans ce contexte pour étudier comment les paramètres du modèle (taux d'absorption, taux d'élimination, etc.) varient d'une strate à une autre, les strates correspondant ici à des groupes de patients définis par le dosage du médicament, le type d'adjuvant, etc. La vraisemblance des modèles non-linéaires à effets mixtes n'ayant typiquement pas de forme explicite, on a généralement recours à des versions stochastiques de l'algorithme EM pour estimer les paramètres de ces modèles, dont SAEM figure parmi les plus utilisés [Delyon et al., 1999].

Nous proposons une extension de SAEM qui permet d'estimer les paramètres des modèles correspondant à plusieurs strates d'observations, en encourageant ces paramètres à être identiques via une pénalité de type fused lasso généralisé. A noter que les similarités sont encouragées tant au niveau des effets fixes que des variances des effets aléatoires. Le fused lasso généralisé est introduit dans l'étape de maximisation de SAEM. En d'autres termes, l'algorithme SAEM pénalisé par une pénalité de type fused lasso généralisé correspond à un SAEM classique, excepté pour l'étape de maximisation. Plus précisément, notre étape de maximisation consiste en une mise à jour des paramètres fixes, des variances des effets aléatoires et des paramètres d'erreur des modèles. Concernant ces derniers paramètres, nous utilisons la mise à jour classique. Pour la mise à jour des effets fixes, le problème d'optimisation correspond à une extension du fused lasso généralisé dans le cas du modèle linéaire où les moindres carrés sont remplacés par des moindres carrés pondérés. Pour la mise à jour des variances des effets aléatoires, nous travaillons sous l'hypothèse, forte, d'indépendance entre les effets aléatoires, si bien que leur matrice de variance-covariance est diagonale. La pénalité porte alors sur les différences entre les éléments diagonaux des matrices de précision (l'inverse des matrices de covariance) et le problème d'optimisation équivaut à une version simplifiée de celui résolu par [Danaher et al., 2014] pour estimer simultanément les struc-

tures de plusieurs modèles graphiques gaussiens. Nous résolvons chacun de ces problèmes d'optimisation via un algorithme de type ADMM (Alternating Direction Method of Multiplier ; voir [Boyd et al., 2011]).

Dans [VV7], nous présentons une étude de simulation où l'on compare notre approche à une stratégie plus classique reposant sur une procédure de sélection de variables pas-à-pas. Cette étude suggère de bonnes performances pour notre approche en matière de sélection de variables sur les configurations considérées. Nous appliquons également notre algorithme sur un jeu de données réel issu de deux essais cliniques en cross-over dans le but d'étudier l'interaction entre le dabigatran etexilate (un anti-coagulant) et trois inhibiteurs de la P-glycoprotéine, en se focalisant sur le paramètre dit de bio-disponibilité. Nous y obtenons des résultats qui semblent pertinents et plausibles aux yeux des experts pharmacologues.

4.2.4 Limites de l'approche : sensibilité au graphe sur des données de grande dimension

Comme établi dans le corollaire 4.2.1 sous des hypothèses assez générales, la version adaptative de CliqueFused renvoie des estimateurs asymptotiquement optimaux si l'on suppose Kp fixe. Dans ce cadre, elle permet également de décrire parfaitement l'effet de Z sur l'association entre Y et $\mathbf{x}$. Elle peut donc être vue comme la méthode de référence dans les situations où la taille de chacune des strates est grande devant le nombre de covariables p.

Cependant, les propriétés de l'approche ne sont pas encore décrites dans le cadre non-asymptotique, et les performances de CliqueFused ne sont donc pas bien connues lorsque certains ratios n_k/p sont petits. Les résultats de [Sharpnack et al., 2012], même s'ils ne traitent pas du cas des données stratifiées et ne concernent que le modèle de suite gaussienne tronquée, suggèrent que lorsque K n'est pas considéré fixe, la version non-adaptative de CliqueFused n'identifie généralement pas correctement les paires $(k_1, k_2) \in [K] \times [K]$ telles que $\beta^*_{k_1,j} = \beta^*_{k_2,j}$ pour $j \in [p]$ fixé, sauf peut-être dans des cas particuliers (homogénéité complète des vecteurs $\boldsymbol{\beta}^*_k$ par exemple). En effet, les cliques utilisées dans l'approche CliqueFused ne sont bien adaptées que lorsqu'il y a peu d'hétérogénéité dans les vecteurs $\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_K$. Notons de plus que les résultats des simulations présentés au chapitre précédent (même s'ils décrivaient le cadre général et non pas la situation spécifique des données stratifiées) allaient dans le sens des résultats théoriques de [Sharpnack et al., 2012].

Outre ses limites théoriques, l'application de CliqueFused se heurte à des problèmes d'ordre pratique. En effet, l'implémentation du fused lasso généralisé n'a été à ce jour effectuée que dans un nombre très restreint de modèles. Dans le logiciel R par exemple, seul le package `GenLasso` [Tibshirani and Taylor, 2011] permet son implémentation, et seulement sous le modèle linéaire (le package FusedLasso [Höfling et al., 2010] est disponible via les archives de R et permet l'implémentation de l'approche sous les modèles linéaire et logistique). Le package gvcm.cat de [Oelker et al., 2014] permet quant à lui l'implémentation d'une version approchée de CliqueFused dans les modèles linéaire, logistique et de Poisson.

En résumé, CliqueFused affiche certaines limites théoriques et pratiques. Je me suis alors intéressé aux propriétés de RefLasso. Cette approche est simple à implémenter (il s'agit d'un lasso simple, sur une transformation des données originales, comme nous le verrons plus précisément dans le paragraphe suivant). Elle peut également être vue comme une version

du fused lasso généralisé, reposant sur un graphe différent de celui utilisé dans CliqueFused. En effet, après avoir choisi la strate de référence $\ell \in [K]$, le critère (4.5) peut se réécrire sous la forme suivante

$$\sum_k \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2} + \lambda_1\|\boldsymbol{\beta}_\ell\|_1 + \lambda_2 \sum_{\substack{k\in[K]\\k\neq\ell}} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_1. \tag{4.9}$$

Deux différences principales existent entre ce critère et celui de CliqueFused, (4.6). Premièrement, seule la norme L_1 de $\boldsymbol{\beta}_\ell$ est pénalisée (et non plus $\sum_{k\in[K]} \|\boldsymbol{\beta}_k\|_1$). Deuxièmement, on ne pénalise pas les $K(K-1)/2$ différences $\|\boldsymbol{\beta}_{k_1} - \boldsymbol{\beta}_{k_2}\|_1$ pour $k_1 < k_2$, mais seulement les $K-1$ différences $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_1$ pour $k \neq \ell$ (ℓ étant fixé). Ainsi, RefLasso peut être vue comme un fused lasso généralisé dont le graphe n'est plus composé de p cliques, mais de p étoiles : pour la j-ème étoile, le centre correspond au paramètre $\beta_{\ell,j}$, chacun des paramètres $\beta_{k,j}$, $k \neq \ell$ étant en périphérie (voire la figure 4.1 pour une illustration). Le graphe sur lequel repose RefLasso est composé de sous-graphes beaucoup moins connectés que dans le cas de CliqueFused, qui peuvent être mieux adaptés à des hétérogénéités parmi les vecteurs $\boldsymbol{\beta}_k^*$. Cependant, la forme de ce graphe implique que RefLasso ne peut que partiellement décrire le rôle de Z sur l'association entre Y et $\mathbf{x}$, puisque les quantités $|\beta_{k_1,j} - \beta_{k_2,j}|$ ne figurent pas dans le terme de pénalité, pour $k_1 \neq \ell$ et $k_2 \neq \ell$.

Alors que CliqueFused semble ne pas être en mesure de décrire le rôle complet de Z (sauf pour sa version adaptative dans un cadre asymptotique en supposant Kp fixe ou peut-être dans des cas où les vecteurs $\boldsymbol{\beta}_k^*$ affichent très peu d'hétérogénéité), on peut se demander si RefLasso fournit une réponse adaptée quant au rôle partiel de Z et permet de détecter les différences entre les vecteurs $\boldsymbol{\beta}_k^*$ et $\boldsymbol{\beta}_\ell^*$, pour le choix ℓ de la strate de référence. Dans le prochain paragraphe, nous étudions les propriétés d'une nouvelle approche, AutoRefLasso, dérivée de RefLasso. AutoRefLasso permet de se défaire du choix arbitraire de la strate de référence a priori et, sous certaines hypothèses, elle identifie automatiquement une strate de référence, *pour chaque covariable*. Nous montrons que sous certaines hypothèses, AutoRefLasso affiche des performances analogues à celle d'ORefLasso, la version oraculaire de RefLasso introduite au paragraphe 4.1, et permet d'étudier le rôle partiel de Z sur l'association entre Y et $\mathbf{x}$ sous des hypothèses typiquement plus faibles que celles requises par RefLasso. D'autre part, le coût algorithmique d'AutoRefLasso est comparable à celui de RefLasso. AutoRefLasso est enfin directement implémentable sous une grande variété de modèles (linéaire, logistique, Poisson, logistique conditionnel, de Cox, etc.) puisque nous montrons que le problème d'optimisation sur lequel repose AutoRefLasso s'écrit lui aussi comme un simple lasso sur une transformation des données originales.

4.3 AutoRefLasso

4.3.1 Principe général

Le point de départ de cette approche consiste à remarquer que la paramétrisation initiale du modèle en (4.1), sur laquelle repose IndepLasso, celle utilisée dans le modèle (4.3) sur la-

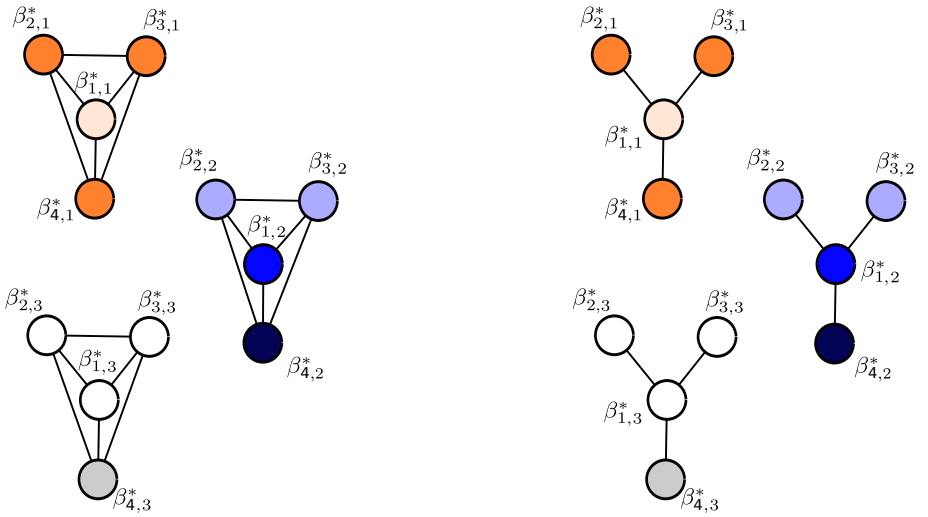


FIGURE 4.1 – Graphes utilisés dans les approches CliqueFused (à gauche) et RefLasso (à droite, avec le choix 1 comme strate de référence) dans le cas $K = 4$ et $p = 3$. CliqueFused correspond à un fused lasso généralisé dont le graphe est constitué de p cliques : un clique par covariables, qui relie l'ensemble des paramètres décrivant l'effet de cette covariable sur les K strates. Pour RefLasso, le graphe est constitué de p graphes en étoile : pour chaque covariable, la strate de référence (ici, la strate 1) est placée au centre de l'étoile, si bien que seules les différences entre les effets sur cette strate de référence et les autres strates sont pénalisés.

quelle repose PoolLasso, et celle utilisée dans l'approche RefLasso sont trois cas particuliers de la paramétrisation suivante,

$$\boldsymbol{\beta}_k^* = \overline{\boldsymbol{\beta}}^* + \boldsymbol{\gamma}_k^*, \quad k \in [K]. \tag{4.10}$$

Cette paramétrisation repose sur $(K+1)p$ paramètres et est donc sur-paramétrée. Ce type de sur-paramétrisation rappelle celle de l'ANOVA où les estimations sont effectuées sous certaines contraintes. Ici, la paramétrisation initiale correspond à la contrainte $\overline{\boldsymbol{\beta}} = \mathbf{0}_p$ alors que la paramétrisation opérée par RefLasso avec le choix ℓ de la strate de référence correspond à la contrainte $\boldsymbol{\gamma}_\ell^* = \mathbf{0}_p$. A noter aussi que parmi l'ensemble des décompositions de la forme (4.10), certaines apparaissent naturellement intéressantes. Le vecteur $\overline{\boldsymbol{\beta}}^*$ peut en effet être vu comme renfermant les p effets « globaux », et les vecteurs $\boldsymbol{\gamma}_k^* \in \mathbb{R}^p$ représenteraient alors les variations des effets sur la strate k autour de ces effets globaux. En ce sens, on pourrait être amené à considérer avec un intérêt particulier les décompositions (4.10) où les composantes du vecteur $\overline{\boldsymbol{\beta}}^*$ sont définies par $\overline{\beta}_j^* = (1/K)\sum_{i=1}^K \beta_{k,j}^*$, $\overline{\beta}_j^* \in \text{médiane}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$, ou encore $\overline{\beta}_j^* \in \text{mode}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$. A noter que ces choix sont équivalents aux définitions suivantes du vecteur $\overline{\boldsymbol{\beta}}^*$,

$$\overline{\boldsymbol{\beta}}^* \in \underset{\overline{\boldsymbol{\beta}} \in \mathbb{R}^p}{\text{argmin}} \sum_{k \in [K]} \|\boldsymbol{\beta}_k^* - \overline{\boldsymbol{\beta}}\|_q,$$

avec $q = 2$, $q = 1$ et $q = 0$ respectivement. Un autre choix intéressant, mais moins intuitif, consiste à définir $\overline{\beta}_j^* \in \text{mode}(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*) = \text{argmin}_{\overline{\beta}_j} \{\mathbb{I}(\overline{\beta}_j \neq 0) + \sum_{k \in [K]} \mathbb{I}(\overline{\beta}_j \neq \beta_{k,j}^*)\}$. Ce choix revient à définir $\overline{\beta}_j^* = \beta_{\ell_j^*}^*$ où ℓ_j^* est la strate supposément renvoyée par l'oracle dans l'approche ORefLasso. La décomposition opérée par la stratégie ORefLasso est donc elle aussi un cas particulier de (4.10). Elle est particulièrement intéressante du point de vue de l'inférence puisqu'elle minimise le nombre de paramètres non nuls à estimer, comme mentionné au paragraphe 4.1.

Suivant le principe de l'approche RefLasso, des estimateurs des paramètres du modèle sur-paramétré (4.10) peuvent être obtenus comme solution minimisant le critère suivant

$$(\widehat{\overline{\boldsymbol{\beta}}}, \widehat{\boldsymbol{\gamma}}_1 \ldots, \widehat{\boldsymbol{\gamma}}_K) \in \underset{\overline{\boldsymbol{\beta}}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K}{\text{argmin}} \left\{ \sum_{k \geq 1} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\overline{\boldsymbol{\beta}} + \boldsymbol{\gamma}_k)\|_2^2}{2} + \lambda_1 \|\overline{\boldsymbol{\beta}}\|_1 + \sum_{k \geq 1} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1 \right\} \tag{4.11}$$

pour des valeurs appropriées des paramètres de régularisation $\lambda_1 \geq 0$ et $\lambda_{2,k} \geq 0$. Travailler avec des valeurs assez élevées de λ_1 équivaut à contraindre $\widehat{\overline{\boldsymbol{\beta}}} = \mathbf{0}_p$, et donc à résoudre les K critères lasso indépendamment, chacun avec le paramètre $\lambda_{2,k}$: IndepLasso est donc un cas particulier de notre approche. D'autre part, travailler avec des valeurs assez élevées de $\lambda_{2,k}$ revient à contraindre $\hat{\boldsymbol{\beta}}_k = \widehat{\overline{\boldsymbol{\beta}}}$ pour tout $k \in [K]$, ce qui correspond à PoolLasso. L'utilisation d'une valeur assez élevée pour $\lambda_{2,\ell}$ équivaut pour sa part à contraindre $\widehat{\boldsymbol{\gamma}}_\ell = \mathbf{0}_p$ (et donc $\widehat{\overline{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}_\ell$) et correspond donc à la stratégie RefLasso avec le choix ℓ pour la strate de référence (ORefLasso est obtenue si le terme $\lambda_{2,k}\|\boldsymbol{\gamma}_k\|_1$ en (4.11) est remplacé par $\sum_{j \in [p]} \lambda_{2,k,j}|\gamma_{k,j}|$ et les valeurs de $\lambda_{2,\ell_j^*,j}$ pour chaque $j \in [p]$ sont assez élevées). Plus généralement, en posant

$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ avec $\tau_k = \lambda_{2,k}/\lambda_1$, et en définissant la version *shrunkée* et $\boldsymbol{\tau}$-pondérée de la médiane de $(b_1, \ldots, b_K)$ comme $\text{WSmedian}(b_1, \ldots, b_K; \boldsymbol{\tau}) = \text{argmin}_b(|b| + \sum_{k\in[K]} \tau_k |b_k - b|)$, il est clair que $\widehat{\overline{\beta}}_j \in \text{WSmedian}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j}; \boldsymbol{\tau})$. En d'autres termes, pour toutes valeurs données des ratios $\tau_k = \lambda_{2,k}/\lambda_1$, notre approche encourage les solutions $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K)$ telles que le vecteur d'effets globaux $\widehat{\overline{\boldsymbol{\beta}}}$ et les vecteurs des différences $\widehat{\boldsymbol{\beta}}_k - \widehat{\overline{\boldsymbol{\beta}}}$ sont creux, avec l'effet global de la j-ème covariable $\widehat{\overline{\beta}}_j$ « identifiée » et définie comme $\text{WSmedian}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j}; \boldsymbol{\tau})$. Dans le contexte où pour tout $j \in [p]$, il existe $\beta_j^* \in \mathbb{R}$ et $K_j^* \subseteq [K]$ tels que $\beta_{k,j}^* = \beta_j^*$ pour tout $j \in K_j^*$, avec $|K_j^{*c}|$ typiquement petit, nous montrons dans le paragraphe 4.3.3 qu'un choix approprié de $\boldsymbol{\tau}$ assure que $\overline{\beta}_j^* = \text{WSmedian}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*; \boldsymbol{\tau}) = \beta_j^*$, ce qui « justifie » la terminologie AutoRefLasso pour cette approche.

4.3.2 Réécriture comme un lasso sur une transformation des données originales

Une propriété intéressante d'AutoRefLasso, RefLasso et ORefLasso est que chacune de ces stratégies peut se réécrire comme un simple lasso sur une transformation des données. Sans perte de généralité, on suppose que $\ell = 1$ est la strate de référence pour la stratégie RefLasso. On suppose de plus qu'un oracle fournit un indice $\ell_j^* \in [K]$ pour chaque $j \in [p]$ pour la stratégie ORefLasso. Soit alors pour tout $k \in [K]$, $P_k = \{j \in [p] : k \neq \ell_j^*\}$ et $\tilde{\mathbf{X}}^{(k)} = \mathbf{X}^{(k)}_{P_k}$. Comme précédemment, on note $\mathscr{Y} = ({\mathbf{Y}^{(1)}}^T, \ldots, {\mathbf{Y}^{(K)}}^T) \in \mathbb{R}^n$ le vecteur contenant les n observations de la variable réponse sur l'ensemble des strates. Alors les critères à minimiser pour les stratégies RefLasso, ORefLasso et AutoRefLasso s'écrivent tous comme

$$\frac{\|\mathscr{Y} - \boldsymbol{\mathcal{X}}\boldsymbol{\theta}\|_2^2}{2} + \lambda_1 \|\boldsymbol{\theta}\|_1, \tag{4.12}$$

où $\boldsymbol{\theta}$ est un vecteur de $\mathbb{R}^{Kp}$ ou $\mathbb{R}^{(K+1)p}$ et $\boldsymbol{\mathcal{X}}$ est l'une des trois matrices suivantes

$$\boldsymbol{\mathcal{X}}^{(1)} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{X}^{(2)} & \frac{\mathbf{X}^{(2)}}{\tau_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \ldots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix}, \quad \widetilde{\boldsymbol{\mathcal{X}}} = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\tilde{\mathbf{X}}^{(1)}}{\tau_1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \ldots & \frac{\tilde{\mathbf{X}}^{(K)}}{\tau_K} \end{pmatrix},$$

$$\bar{\boldsymbol{\mathcal{X}}} = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \ldots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix},$$

pour des valeurs données $\tau_k > 0$, $k \in [K]$. Pour AutoRefLasso, les solutions $\hat{\bar{\boldsymbol{\theta}}} \in \mathbb{R}^{(K+1)p}$ de (4.12) avec $\boldsymbol{\mathcal{X}} = \bar{\boldsymbol{\mathcal{X}}}$ fournissent des estimateurs de $\bar{\boldsymbol{\theta}}^* = (\overline{\boldsymbol{\beta}}^{*T}, \tau_1 {\boldsymbol{\gamma}_1^*}^T, \ldots, \tau_K {\boldsymbol{\gamma}^*}^T)^T$, avec $\overline{\beta}_j^* = \text{WSmedian}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*; \boldsymbol{\tau})$ et $\boldsymbol{\gamma}_k^* = \boldsymbol{\beta}_k^* - \overline{\boldsymbol{\beta}}^*$. Pour RefLasso, les solutions de (4.12) avec $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}^{(1)}$ fournissent des estimateurs de $\boldsymbol{\theta}_1^* = ({\boldsymbol{\beta}_1^*}^T, \tau_1 {\boldsymbol{\delta}_2^*}^T, \ldots, \tau_K {\boldsymbol{\delta}^*}^T)^T \in \mathbb{R}^{Kp}$, avec $\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$. Finalement, pour ORefLasso, les solutions $\hat{\tilde{\boldsymbol{\theta}}} \in \mathbb{R}^{Kp}$ de (4.12) avec $\boldsymbol{\mathcal{X}} = \tilde{\boldsymbol{\mathcal{X}}}$ sont des estimateurs de $\tilde{\boldsymbol{\theta}}^* = ({\tilde{\boldsymbol{\beta}}^*}^T, \tau_1 {\tilde{\boldsymbol{\delta}}_1^*}^T, \ldots, \tau_K {\tilde{\boldsymbol{\delta}}^*}^T)^T$ avec $\tilde{\beta}_j^* = \beta^*_{\ell_j^*, j}$ et $\tilde{\boldsymbol{\delta}}_k^* = (\boldsymbol{\beta}_k^* - \tilde{\boldsymbol{\beta}}^*)_{P_k}$.

Cette propriété de réécriture comme un lasso s'étend naturellement à l'ensemble des modèles linéaires généralisés, aux modèles de Cox, etc. Sous des modèles de régression logistique par exemple, les critères relatifs à RefLasso, ORefLasso et AutoRefLasso s'écrivent comme des lasso dans le cas logistique,

$$-\mathcal{L}_{\text{logistic}}(\mathscr{Y}, \boldsymbol{\mathcal{X}}\boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{\theta}\|_1,$$

avec $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\theta}$ définis comme dans le cas linéaire et, pour tout $\mathbf{y} \in \{0,1\}^n$ et $\mathbf{z} \in \mathbb{R}^n$, $\mathcal{L}_{\text{logistic}}(\mathbf{y}, \mathbf{z}) = \sum_{i\in[n]}\{y_i z_i - \log(1+e^{z_i})\}$. Cette propriété est particulièrement intéressante pour notre stratégie AutoRefLasso puisqu'elle la rend directement implémentable sous une large variété de modèles de régressions, en fait tous ceux pour lesquels le lasso a été implémenté. Le package `glmnet` de R [Friedman et al., 2010] permet ainsi de traiter les modèles linéaire, logistique, de Poisson, de Cox, etc. (à noter également que `glmnet` peut tirer profit de la structure creuse de la matrice $\bar{\boldsymbol{\mathcal{X}}}$, en particulier lorsque Kp est grand, pour améliorer les temps de calcul). Plus généralement, cette propriété établit qu'il n'y a pratiquement pas de surcoût computationel lié à l'utilisation d'AutoRefLasso par rapport à la stratégie RefLasso.

4.3.3 Sélection de variables dans un cadre non-asymptotique

La réécriture (4.12) n'est pas seulement intéressante du point de vue de l'implémentation mais aussi pour étudier les propriétés théoriques d'AutoRefLasso, et en particulier pour les comparer à celles de RefLasso et ORefLasso. Dans ce paragraphe, nous étudions la *sparsistency* (consistance en sélection de variables) de ces approches. Pour que le lasso soit sparsistent, il est maintenant établi que la matrice de design « doit » vérifier la condition d'irreprésentabilité, cette condition étant suffisante et « presque nécessaire » [Zhao and Yu, 2006, Wainwright, 2009]. Avec la formulation (4.12) du lasso et en notant $\boldsymbol{\theta}^*$ le vecteur de paramètre théorique et J^* son support, la matrice $\boldsymbol{\mathcal{X}}$ vérifie la condition d'irreprésentabilité si et seulement si $\Lambda_{\min}(\boldsymbol{\mathcal{X}}_{J^*}^T \boldsymbol{\mathcal{X}}_{J^*}) \geq C_{\min}$ pour une valeur fixée $C_{\min} > 0$ et

$$\max_{j \notin J^*} \|(\boldsymbol{\mathcal{X}}_{J^*}^T \boldsymbol{\mathcal{X}}_{J^*})^{-1} \boldsymbol{\mathcal{X}}_{J^*}^T \mathcal{X}_j\|_1 < 1,$$

avec $\mathcal{X}_j$ la j-ème colonne de $\boldsymbol{\mathcal{X}}$. Autrement dit, la condition d'irreprésentabilité assure que le modèle restreint à J^* est identifiable et que les colonnes de J^{*c} ne sont pas trop alignées sur celles de J^*. Dans ce paragraphe, nous établissons des conditions, notamment sur les paramètres τ_k, assurant que $\boldsymbol{\mathcal{X}}^{(1)}$, $\tilde{\boldsymbol{\mathcal{X}}}$ et $\bar{\boldsymbol{\mathcal{X}}}$ vérifient la condition d'irreprésentabilité de telle sorte que RefLasso, ORefLasso et AutoRefLasso puissent être sparsistent, à condition d'être utilisés avec une valeur appropriée de λ_1 et que le signal soit assez élevé (condition de type "beta-min").

Même si des cas plus généraux peuvent être traités (voir le Supplementary Material de [VV8]), nous nous concentrons ici sur le cas simple suivant, par souci de simplification des notations et de l'interprétation notamment. Nous supposons que les strates sont équilibrées et que les designs sont orthogonaux dans chaque strate ; plus précisément, nous supposons que $n_k = n/K$ et $(\mathbf{X}^{(k)^T}\mathbf{X}^{(k)})/n_k = \mathbf{I}_{n_k}$ pour tout $k \in [K]$. On supposera de plus que pour chaque $j \in [p]$ il existe un mode unique $\beta_j^* \in \mathbb{R}$ de l'ensemble $(0, \beta_{1,j}^*, \ldots, \beta_{K,j}^*)$, et

on définit $K_j^* = \{k \in [K] : \beta_{k,j}^* = \beta_j^*\}$. Pour la stratégie ORefLasso, on supposera qu'un oracle renvoie un indice $\ell_j^* \in K_j^*$ pour chaque $j \in [p]$. Finalement, on supposera que pour tout $k \in [K]$, $\tau_k = \tau/\sqrt{K}$ pour une valeur $\tau > 0$ donnée, et que $n_k^{-1/2}\|X_j^{(k)}\|_2 \leq 1$ pour chaque $(k,j) \in [K] \times [p]$. Combinées, ces hypothèses assurent que les colonnes de $\boldsymbol{\mathcal{X}}$ sont de norme L_2 comparables, avec $\boldsymbol{\mathcal{X}}$ égal à $\boldsymbol{\mathcal{X}}^{(1)}$, $\bar{\boldsymbol{\mathcal{X}}}$ ou $\tilde{\boldsymbol{\mathcal{X}}}$. Plus précisément, en notant $\mathcal{X}_j$ la j-ème colonne de $\boldsymbol{\mathcal{X}}^{(1)}$, $\bar{\boldsymbol{\mathcal{X}}}$ ou $\tilde{\boldsymbol{\mathcal{X}}}$, elles assurent que $n^{-1}\|\mathcal{X}_j\|_2 \leq \max(1, \tau^{-1})$ pour tout $j \in [(K+1)p]$.

Dans ce contexte, l'objectif principal relatif à la sélection de variables est de retrouver les ensembles $S^* = \{j \in [p] : \beta_j^* \neq 0\}$ et $T^* = \{(k,j) \in [K] \times [p] : \beta_{k,j}^* \neq \beta_j^*\}$, c'est-à-dire le sous-ensemble des covariables dont l'effet global est non-nul et le sous-ensemble des paires strate/covariable où l'on observe des hétérogénéités (i.e. où l'effet de la covariable sur la strate est différent de son effet global). Notons que l'hypothèse d'unicité du mode implique notamment que $\min_{j \in S^*} |K_j^*| > 1$.

Sous les différentes hypothèses mentionnées ci-dessus, on obtient premièrement les deux lemmes suivants, dont les preuves figurent dans le Supplementary Material de [VV8].

Lemme 4.3.1 *Les matrices $\bar{\boldsymbol{\mathcal{X}}}$ et $\tilde{\boldsymbol{\mathcal{X}}}$ vérifient la condition d'irreprésentabilité si et seulement si*

$$(\mathbf{IC}) \qquad 0 \leq \frac{\sqrt{K}}{K - 2\mathcal{D}_1} < \tau < \frac{\sqrt{K}}{\mathcal{D}_0},$$

avec $\mathcal{D}_0 = \max_{j \notin S^*} |K_j^{*c}|$ si $S^* \neq [K]$ et 0 sinon, et $\mathcal{D}_1 = \max_{j \in S^*} |K_j^{*c}|$ si $S^* \neq \emptyset$ et $-\infty$ sinon.

Soit $S^{(1)*} = \{j \in [p] : \beta_{1,j}^* \neq 0\}$. La matrice $\boldsymbol{\mathcal{X}}^{(1)}$ vérifie la condition d'irreprésentabilité si et seulement si

$$(\mathbf{IC}^{(1)}) \qquad 0 \leq \frac{\sqrt{K}}{K - 2\mathcal{D}_1^{(1)}} < \tau < \frac{\sqrt{K}}{\mathcal{D}_0^{(1)}},$$

avec $\mathcal{D}_0^{(1)} = \max_{j \notin S^{(1)*}} |\{k \in [K] : \beta_{k,j}^* \neq \beta_{1,j}^*\}|$ si $S^{(1)*} \neq [K]$ et 0 sinon, et $\mathcal{D}_1^{(1)} = \max_{j \in S^{(1)*}} |\{k \in [K] : \beta_{k,j}^* \neq \beta_{1,j}^*\}|$ si $S^{(1)*} \neq \emptyset$ et $-\infty$ sinon.

Lemme 4.3.2 *Sous la condition* (**IC**)*, on a $\overline{\beta}_j^* = \mathrm{WSmedian}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*; \boldsymbol{\tau}) = \beta_j^* = \beta_{\ell_j^*,j}^*$.*

Notons tout d'abord que sous (**IC**), on a forcément $2\mathcal{D}_1 + \mathcal{D}_0 < K$. De manière analogue, sous ($\mathbf{IC}^{(1)}$) on a forcément $2\mathcal{D}_1^{(1)} + \mathcal{D}_0^{(1)} < K$. Le Lemme 4.3.1 établit que les matrices $\bar{\boldsymbol{\mathcal{X}}}$ et $\tilde{\boldsymbol{\mathcal{X}}}$ des stratégies AutoRefLasso et ORefLasso, respectivement, vérifient la condition d'irreprésentabilité sous la même condition sur τ. Sous cette condition, le Lemme 4.3.2 établit par ailleurs que $\bar{\boldsymbol{\theta}}_{\bar{J}^*}^* = \tilde{\boldsymbol{\theta}}_{\tilde{J}^*}^*$, avec $\bar{J}^* = \mathrm{supp}(\bar{\boldsymbol{\theta}}^*)$ et $\tilde{J}^* = \mathrm{supp}(\tilde{\boldsymbol{\theta}}^*)$ (pour rappel, les définitions de $\bar{\boldsymbol{\theta}}^*$ et $\tilde{\boldsymbol{\theta}}^*$ sont données tout de suite après l'Equation (4.12)). Comme nous l'établissons plus précisément dans le Théorème 4.3.1 ci-dessous, cela implique qu'AutoRefLasso permet d'identifier S^* et T^* sous (approximativement) les mêmes hypothèses que celles requises par ORefLasso, sans avoir à imposer que les ℓ_j^* soient connus par avance. D'autre part, si $\{1\} \in \cap_{j \in [p]} K_j^*$ alors (**IC**) et ($\mathbf{IC}^{(1)}$) sont identiques (et RefLasso revient alors à ORefLasso). Par contre, si $\{1\} \notin \cap_{j \in [p]} K_j^*$, non seulement $T^{(1)*} \neq T^*$ avec

$T^{(1)*} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{1,j}\}$ (et potentiellement $S^{(1)*} \neq S^*$), mais ($\mathbf{IC}^{(1)}$) est également généralement plus forte que (**IC**). En d'autres termes, RefLasso est moins souvent capable d'identifier $S^{(1)*}$ et $T^{(1)*}$ avec grande probabilité, que ne le sont ORefLasso et AutoRefLasso d'identifier S^* et T^*.

Remarque 4.3.1 *Le cadre considéré ici est simpliste (designs orthogonaux dans chaque strate, équilibrée), et peu réaliste en pratique (il couvre néanmoins l'ANOVA à un facteur et le modèle tronqué de suites gaussiennes). Il est cependant utile puisqu'il donne un éclairage sur le type d'hypothèses « nécessaires » pour la consistance en sélection de variable du lasso dans un cas particulier de modèle incluant des interactions. L'approche RefLasso peut en effet être vue comme modélisant les interactions entre la variable Z, ici catégorielle et incluse dans le modèle via des variables indicatrices, et le vecteur $\mathbf{x}$ (les interactions étant incluses dans le modèle via des produits). Le cadre considéré ici permet d'expliciter ce qu'induit, dans ce cadre simple, la condition* $\max_{j \notin J^*_1} \|(\boldsymbol{\mathcal{X}}^{(1)T}_{J^*_1} \boldsymbol{\mathcal{X}}^{(1)}_{J^*1})^{-1} \boldsymbol{\mathcal{X}}^{(1)T}_{J^*_1} \mathcal{X}^{(1)}_j\|_1 < 1$, *où J^*_1 est le support du vecteur $\boldsymbol{\theta}^*_1 = (\boldsymbol{\beta}^{*T}_1, \tau_1 \boldsymbol{\delta}^{*T}_2, \ldots, \tau_K \boldsymbol{\delta}^{*T})^T \in \mathbb{R}^{Kp}$. Cette condition devient ici $2\mathcal{D}^{(1)}_1 + \mathcal{D}^{(1)}_0 < K$. Elle stipule donc que, pour chaque covariable, son effet sur la plupart des strates est égal à celui sur la strate de référence, choisie a priori. La condition que doit vérifier la version oraculaire de RefLasso, ORefLasso, ainsi que notre approche AutoRefLasso, reste forte. Mais elle est généralement moins forte que celle que doit vérifier RefLasso puisqu'elle stipule « seulement » que pour chaque covariable, son effet sur la plupart des strates vaut* $\beta^*_j = \text{mode}(0, \beta^*_{1,j}, \ldots, \beta^*_{K,j})$.

En se concentrant maintenant sur AutoRefLasso et ORefLasso, on peut établir le résultat suivant qui décrit des conditions sous lesquelles S^* et T^* sont identifiés avec grande probabilité.

Théorème 4.3.1 *Pour tout $k \in [K]$, supposons que les $\varepsilon^{(k)}_i$, $i \in [n_k]$, sont des variables i.i.d. sous-gaussiennes centrées, de paramètre $\sigma > 0$. Sous l'hypothèse* (**IC**)*, soit alors*

$$\gamma = \min\left(1 - \frac{\mathcal{D}_0\tau}{\sqrt{K}}, 1 - \frac{\sqrt{K} + \mathcal{D}_1\tau}{(K - \mathcal{D}_1)\tau}\right)$$

et

$$C_{\min} = \min\left(1, \frac{1}{\tau^2}, \frac{1}{2}\left[\left(\frac{1}{\tau^2} + 1\right) - \sqrt{\left(\frac{1}{\tau^2} - 1\right)^2 + \frac{4\mathcal{D}_1}{\tau^2 K}}\right]\right).$$

Pour $\eta \in \{0,1\}$, on définit

$$\lambda^{(\eta)}_1 > \frac{2}{\gamma \min(1,\tau)} \sqrt{2\sigma^2 n \log((K+\eta)p)} \quad \text{et} \quad \lambda^{(\eta)}_{2,k} = \tau_k \lambda^{(\eta)}_1$$

et on introduit

$$\beta^{(\eta)}_{\min} = \frac{\lambda^{(\eta)}_1}{n}\left(\frac{\sqrt{|S^*| + |T^*|}}{C_{\min}} + \frac{4\sigma}{\sqrt{C_{\min}}}\right).$$

Finalement, considérons les conditions de type β-min suivantes :

$$(\mathbf{C}_{\beta^{(\eta)}_{\min}})(i):\ \forall j\in S^*,\ |\beta^*_j|>\beta^{(\eta)}_{\min};\quad (\mathbf{C}_{\beta^{(\eta)}_{\min}})(ii):\ \forall j\in[p],\forall k\notin K^*_j,\ |\beta^*_{k,j}-\beta^*_j|>\frac{\sqrt{K}\beta^{(\eta)}_{\min}}{\tau}.$$

Alors, S^* et T^* sont tous deux identifiés

- *avec une probabilité supérieure à $1-4\exp(-c_1{\lambda^{(0)}_1}^2)$, pour une constante $c_1>0$, par ORefLasso lancé avec les paramètres $\lambda_1=\lambda^{(0)}_1$ et $\lambda_{2,k}=\lambda^{(0)}_{2,k}$ sous $(\mathbf{C}_{\beta^{(0)}_{\min}})(i-ii)$, et on a $\|\hat{\tilde{\boldsymbol{\theta}}}_{\tilde{J}^*}-\tilde{\boldsymbol{\theta}}^*_{\tilde{J}^*}\|_\infty\leq\beta^{(0)}_{\min}$;*
- *avec une probabilité supérieure à $1-4\exp(-c_1{\lambda^{(1)}_1}^2)$, pour une constante $c_1>0$, par AutoRefLasso lancé avec les paramètres $\lambda_1=\lambda^{(1)}_1$ et $\lambda_{2,k}=\lambda^{(1)}_{2,k}$ sous $(\mathbf{C}_{\beta^{(1)}_{\min}})(i-ii)$, et on a $\|\hat{\bar{\boldsymbol{\theta}}}_{\bar{J}^*}-\bar{\boldsymbol{\theta}}^*_{\bar{J}^*}\|_\infty\leq\beta^{(1)}_{\min}$.*

Ce résultat s'obtient à partir du Théorème 1 de [Wainwright, 2009] ; une hypothèse implicite est que K et/ou p diverge avec n si bien que $1-4\exp(-c_1{\lambda^{(\eta)}_1}^2)\to 1$ lorsque $n\to\infty$. Si $\max_{j\in[p]}|K^{*c}_j|<K/3$, le Théorème 4.3.1 montre clairement que, dans le cas équilibré et orthogonal, AutoRefLasso est capable d'identifier S^* et T^* avec grande probabilité sous des conditions analogues à celles que requerrait ORefLasso, sans pour autant avoir à supposer que les ℓ^*_j sont donnés par avance. Dans le Supplementary Material de [VV8], nous montrons comment se résultat s'étend au cas de strates non équilibrées et/ou à des designs non orthogonaux. Dans le cas le plus général, les conditions assurant l'identification de S^* et T^* avec grande probabilité sont un peu plus fortes pour AutoRefLasso que pour ORefLasso.

Une autre remarque concerne la valeur de $\beta^{(\eta)}_{\min}$. Pour faciliter l'interprétation, considérons les cas où $\mathcal{D}_0=\mathcal{D}_1=\mathcal{D}$ dans un cadre asymptotique où K (et potentiellement p) diverge(nt) avec n, tout comme $|T^*|$ (et potentiellement $|S^*|$). Si $\mathcal{D}\ll\sqrt{K}$ ou $\mathcal{D}=c\sqrt{K}$ pour une constante $0<c\leq 1/2$, alors le choix $\tau=1$ assure l'identification des supports pour des signaux tels que $\beta^{(\eta)}_{\min}=\mathcal{O}(\sqrt{(|S^*|+|T^*|)\log((K+1)p)/n})$, ce qui est optimal au terme logarithmique près. Si $\mathcal{D}=c\sqrt{K}$ pour une constante $c>1/2$, on obtient le même ordre de grandeur pour $\beta^{(\eta)}_{\min}$ mais pour le choix $\tau=(2c)^{-1}<1$. Par contre, si $\mathcal{D}\gg\sqrt{K}$, alors on observe un changement de régime. Pour $\sqrt{K}\ll\mathcal{D}\ll K$ le choix optimal est $\tau=\sqrt{K}/(2\mathcal{D})$ qui n'assure l'identification des supports que si $\beta^{(\eta)}_{\min}=\mathcal{O}((\mathcal{D}/\sqrt{K})\times\sqrt{(|S^*|+|T^*|)\log((K+1)p)/n})$. Finalement, si $\mathcal{D}=cK$ pour une constante $0<c<1/3$, alors le résultat du Théorème 4.3.1 est pratiquement vide de sens : le choix optimal pour τ est $\mathcal{O}(1/\sqrt{K})$ qui n'assure l'identification des supports que si $\beta^{(\eta)}_{\min}=\mathcal{O}(\sqrt{K(|S^*|+|T^*|)\log((K+1)p)/n})$. En voyant ORefLasso comme une version du fused lasso généralisé reposant sur un graphe composé de p sous-graphes en étoile, ces résultats suggèrent une nouvelle fois que le graphe du fused lasso généralisé doit être en assez bonne adéquation avec la véritable structure du vecteur à estimer pour assurer la sparsistency de l'approche : dans notre cas, il apparaît que le nombre de paramètres différents de β^*_j doit être au plus de l'ordre de $\sqrt{K}$ pour assurer la sparsistency d'ORefLasso (et AutoRefLasso) au niveau de signal optimal $\beta^{(\eta)}_{\min}=\mathcal{O}(\sqrt{(|S^*|+|T^*|)\log((K+1)p)/n})$.

Une dernière remarque est que, comme attendu, il est plus difficile d'identifier T^* que S^*, au sens où les hétérogénéités doivent être au moins de magnitude $|\beta^*_{k,j} - \beta^*_j| > \sqrt{K}\beta^{(\eta)}_{\min}/\tau$ pour $k \notin K^*_j$ pour être retrouvées, alors que les composantes non-nulles $|\beta^*_j|$ doivent seulement être supérieures à $\beta^{(\eta)}_{\min}$. L'identification de T^* est encore plus difficile dans le cas de strates non équilibrées (les hétérogénéités sur les strates de faible effectif étant les plus délicates à identifier).

4.3.4 Illustrations

Dans [VV8], nous illustrons sur données simulées les performances d'AutoRefLasso, RefLasso, ORefLasso et CliqueFused. L'objectif principal est de compléter nos résultats théoriques. Sous des designs orthogonaux dans chaque strate et équilibrés, ceux-ci établissent notamment l'existence de valeurs des paramètres de régularisation λ_1 et λ_2 telles que les ensemble S^* et T^* sont identifiés par ORefLasso et AutoRefLasso avec grande probabilité, si les vecteurs $\boldsymbol{\beta}^*_k$ sont assez homogènes. Dans notre étude de simulation, on a alors cherché à évaluer les performances d'AutoRefLasso notamment sous des designs non orthogonaux, et pour des choix des paramètres λ_1 et τ reposant sur les données. Nos résultats empiriques confirment qu'AutoRefLasso et ORefLasso partagent des performances analogues, et sont généralement supérieurs à RefLasso et CliqueFused, sous les designs considérés. Ils confirment également que pour les performances sont dégradées lorsque le niveau d'homogénéité entre les vecteurs $\boldsymbol{\beta}^*_k$ augmente.

Nous illustrons également les approches AutoRefLasso, RefLasso et CliqueFused sur un jeu de données de « cellules uniques » décrivant les niveaux d'expressions de 45 facteurs de transcriptions dans les cellules, à huit instants après le déclenchement de la différentiation des cellules ($H0, H1, H6, H12, H24, H48, H72$ et $H96$). Pour chaque instant, qui définit ici une strate, les données relatives à $n_k = 120$ cellules sont disponibles, $k = 1, \ldots, 8$. Ce jeu de données est décrit dans [Kouno et al., 2013], où les auteurs proposent d'étudier les variations parmi les associations entre les facteurs de régulation au cours du temps. Leur approche est basique et repose sur les corrélations, alors que le recours aux modèle graphiques gaussiens semblerait plus judicieux. Ici, nous nous contentons d'étudier les variations des associations entre un facteur de transcription donné, EGR2, et les $p = 44$ autres facteurs, sous un modèle de régression linéaire. Nous considérons les approches AutoRefLasso et CliqueFused, ainsi que RefLasso, avec les choix de strate de référence $H0$ et $H96$. Les paramètres de régularisation sont sélectionnés par validation croisée, dans ce contexte où n_k/p est relativement faible. Les estimations des vecteurs de paramètres $\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_8$ retournées par chacune des approches sur les 8 strates horaires sont représentées sur la figure 4.2. Même si on ne connait bien sûr pas la vérité sur ce jeu de données, ces résultats illustrent que CliqueFused détecte beaucoup moins d'hétérogénéités sur ces données (une seule hétérogénéité est détectée, pour la variable MYB en H0). Ils illustrent également l'impact du choix de la strate de référence dans l'approche RefLasso : pour certaines covariables, les « profils » d'association avec EGR2 au cours du temps sont très différents en fonction du choix de la strate de référence. Par exemple, reprenant l'exemple de MYB, AutoRefLasso, CliqueFused et RefLasso avec le choix $H96$ pour la strate de référence renvoient des profils suggérant que l'association avec EGR2 est constante entre $H1$ et $H96$, mais moins marquée

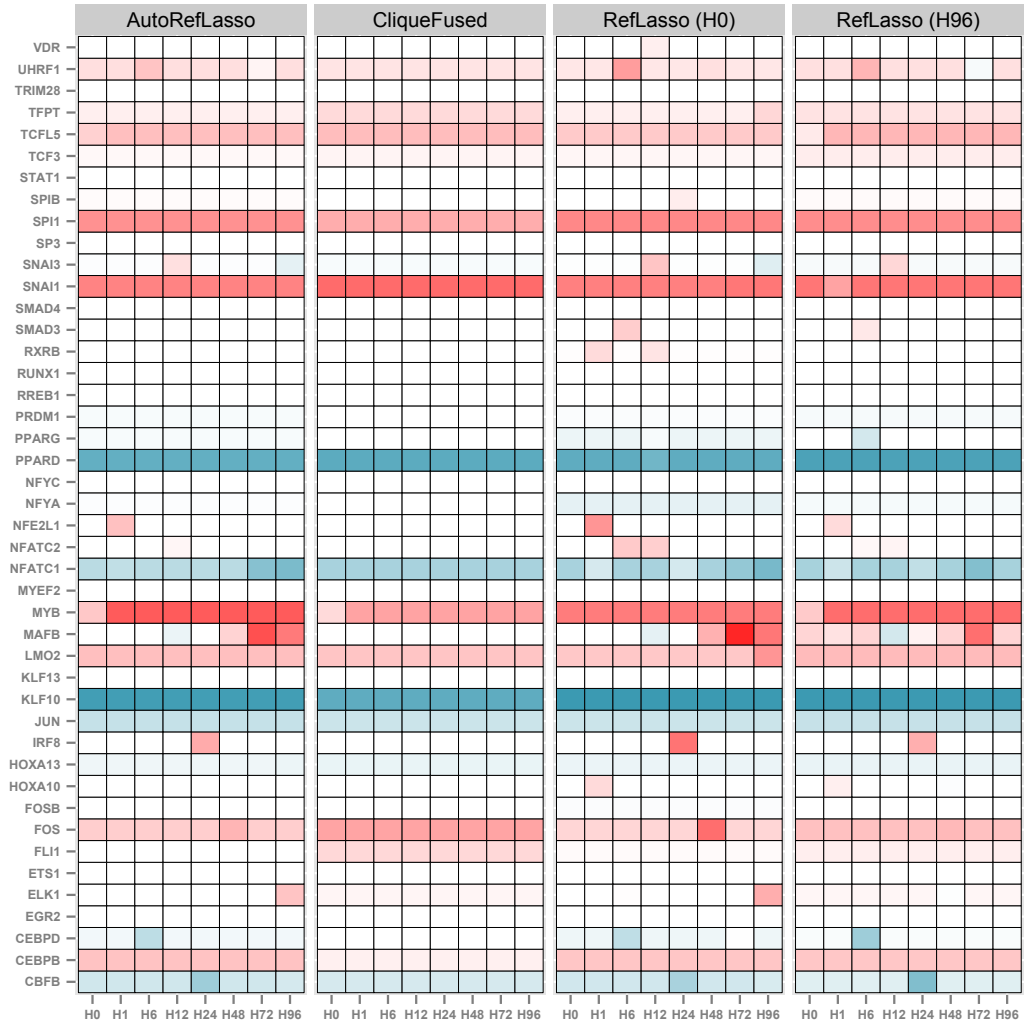


FIGURE 4.2 – Estimation des paramètres du modèle linéaire pour le facteur de transcription EGR2 dans 8 strates horaires. Quatre approches sont considérées : AutoRefLasso, CliqueFused, et RefLasso pour les choix de la strate de référence $H0$ et $H96$.

en $H0$. RefLasso avec le choix $H0$ comme strate de référence ne détecte quant à lui aucune hétérogénéité. Dans le cas de ELK1, AutoRefLasso et RefLasso avec le choix $H0$ comme strate de référence suggèrent un profil constant entre $H0$ et $H72$ (aucune association entre ELK1 et EGR2 à ces instants là), et une association positive en $H96$. RefLasso avec le choix $H96$ suggère un profil bien différent. Mêmes si elles doivent être interprétées avec précaution, nous avons calculé les pvalues des tests de Wald après estimation par MCO sous les modèles identifés par chaque approche. Considérant le modèle retourné par AutoRefLasso, l'hétérogénéité en $H0$ pour MYB est significative, de même que celle détectée en $H96$ pour ELK1. Considérant le modèle retourné par RefLasso avec la strate $H96$ comme référence par exemple, l'effet commun détecté sur $H0, H1, H6, H12, H24, H72$ et $H96$ n'est pas significatif, pas plus que l'hétérogénéité détectée en $H48$. Du point de vue du pouvoir prédictif, nous avons évalué par validation croisée l'erreur de prédiction de chacune des quatre approches et AutoRefLasso affiche les meilleures performances prédictives, et CliqueFused les plus modestes. Ainsi, sur cet exemple, AutoRefLasso semble être le plus à même de décrire les hétérogénéités parmi les vecteurs $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_8^*$ et retourne en tout cas des estimations présentant le meilleur pouvoir prédictif.

4.4 Projet

Une part importante de mon projet de recherche à moyen terme concerne l'étude de méthodes adaptées au cas des données stratifiées, et leur application en épidémiologie. Deux de ces projets sont décrits dans les paragraphes suivants (un autre projet sera décrit dans le chapitre suivant, qui couvre l'estimation de la structure des modèles graphiques binaires).

4.4.1 Approfondissements autour d'AutoRefLasso

Un premier projet concerne diverses extensions autour d'AutoRefLasso, et des comparaisons approfondies, notamment avec CliqueFused.

Dans le cadre asymptotique, et en supposant p fixe, CliqueFused apparaît comme la méthode de référence, et est en tout cas préférable à RefLasso ou AutoRefLasso. Elle seule permet l'étude complète du rôle de Z sur l'association entre Y et $\mathbf{x}$ et l'identification de plusieurs groupes de strates sur lesquelles l'effet d'une variable est constant : avec RefLasso ou AutoRefLasso, on ne peut espérer identifier qu'un groupe de strates sur lesquelles l'effet est constant, les effets estimés sur les autres étant tous différents par construction. Une question qui me semble intéressante en pratique est la suivante : une utilisation itérée d'AutoRefLasso permettrait-elle de détecter plusieurs groupes de paramètres égaux ? Dans le cas où Kp est fixe, il est aisé de montrer qu'une version adaptative d'AutoRefLasso, appliquée itérativement selon un shéma adapté, détecterait en effet ces groupes avec probabilité tendant vers un lorsque $n \to \infty$ (en utilisant les propriétés oraculaires du lasso adaptatif de [Zou, 2006] par exemple). L'étude de cette stratégie itérative dans un cadre non-asymptotique pourrait permettre d'obtenir des hypothèses assurant l'identification de plusieurs groupes de strates sur lesquelles l'effet d'une covariable donnée est constant, sans supposer Kp fixe.

Un autre point qui mérite quelques éclaircissements concerne le cas où les strates ne sont pas équilibrées. Concernant la version adaptative de CliqueFused, les résultats asymptotiques (dans le cas où Kp est fixe) discutés dans ce document, ainsi que ceux établis dans [Gertheiss and Tutz, 2012, Oelker et al., 2014], reposent sur l'hypothèse selon laquelle les strates ont des tailles tendant vers l'infini à la même vitesse. Concernant AutoRefLasso, nos résultats sont établis pour le choix $\tau_k = \lambda_{2,k}/\lambda_1 = \tau\sqrt{n_k/n}$, pour $k \in [K]$, qui assure que les colonnes de $\bar{\boldsymbol{\mathcal{X}}}$ sont normalisées dès lors que les colonnes de chacune des matrices de design $\mathbf{X}^{(k)}$ le sont. Sous les hypothèses classiques pour établir les propriétés non-asymptotiques des estimateurs lasso, le fait de travailler avec des colonnes normalisées améliore ses propriétés : on divise par exemple l'erreur d'estimation $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$ par le facteur $(\max_{j\in J_0(\boldsymbol{\theta}^*)} C_j)/(\max_{j\in[d]} C_j)$ en normalisant les colonnes de la matrice de design $\boldsymbol{\mathcal{X}}$, avec $C_j = \|\mathcal{X}_j\|_2/\sqrt{n}$, $J_0(\boldsymbol{\theta}^*) = \{j \in [d] : \theta_j^* \neq 0\}$ et $\boldsymbol{\theta}^* \in \mathbb{R}^d$ le paramètre du modèle (la matrice $\boldsymbol{\mathcal{X}}$ étant de dimension $n \times d$). Dans le cas d'AutoRefLasso, la normalisation induite par ce choix pour les ratios τ_k impliquerait ainsi de bonnes propriétés pour l'erreur d'estimation $\|\hat{\bar{\boldsymbol{\theta}}} - \bar{\boldsymbol{\theta}}^*\|_2^2$, avec en particulier $\bar{\boldsymbol{\theta}}^* = (\overline{\boldsymbol{\beta}}^{*T}, \tau_1\boldsymbol{\gamma}_1^{*T}, \ldots, \tau_K\boldsymbol{\gamma}^{*T})^T \in \mathbb{R}^{(K+1)p}$. Il pourrait être intéressant d'étudier le comportement de $\sum_{k\in[K]} \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_2^2$ en fonction de ce choix (avec en particulier $\boldsymbol{\beta}_k^* = \overline{\boldsymbol{\beta}} + \boldsymbol{\gamma}_k^*$).

Le choix $\tau_k = \tau\sqrt{n_k/n}$ « tire » naturellement l'effet global estimé de chaque variable $\widehat{\overline{\beta}}_j$ vers les effets estimés sur les strates de plus grands effectifs (puisqu'à l'optimum, on a $\widehat{\overline{\beta}}_j = \mathrm{WSmedian}(\hat{\beta}_{1,\mathrm{j}}, \ldots, \hat{\beta}_{\mathrm{K,j}}; \boldsymbol{\tau})$) et privilégie ainsi sans doute une bonne estimation des paramètres sur ces strates. Ce phénomène est accentué par le fait que l'adéquation aux données est mesurée par le terme $\sum_{k\in[K]} \|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2/2$: chaque observation a le même poids mais globalement, les observations des strates de grand effectif pèsent plus que les autres. Un autre critère, « rééquilibrant » le poids de chaque strate, pourrait être défini en remplaçant ce terme d'accroche aux données par

$$\sum_{k\in[K]} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{n_k}.$$

Le nouveau critère qui en résulte correspond toujours à un lasso, mais où les moindres carrés sont remplacés par des moindres carrés pondérés, les observations de la strate $k \in [K]$ ayant un poids $1/n_k$. Il est intéressant de noter qu'on a toujours $\widehat{\overline{\beta}}_j = \mathrm{WSmedian}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j}, \boldsymbol{\tau})$ à l'optimum, mais bien sûr les $\hat{\boldsymbol{\beta}}_k$ sont différents, et se « focalisent » plus sur les strates de petits effectifs. Il serait intéressant d'étudier les propriétés des estimateurs ainsi obtenus, notamment en matière d'erreur d'estimation et de prédiction.

D'autre part, dans l'optique d'une application en épidémiologie ou en recherche clinique, une question qui se pose est celle de la significativité mesurée par la p-value, ou celle de la précision des estimations (décrite par les intervalles de confiance). Au vu de la réécriture d'AutoRefLasso sous forme d'un simple lasso, on peut espérer pouvoir adapter certaines des approches proposées dans la littératures de l'inférence post-sélection pour les estimateurs lasso, en particulier les *knockoffs* de [Barber and Candès, 2015], les projections régularisées de [Zhang and Zhang, 2014] (voir aussi [Van de Geer et al., 2014]), les approches de ré-échantillonnage [Dezeure et al., 2014] (voir aussi [Meinshausen et al., 2009]).

L'adaptation de ces approches n'est cependant pas complètement triviale du fait de la surparamétrisation sur laquelle repose AutoRefLasso et les problèmes d'identifiabilité inhérents à notre approche, et plus généralement à ces données stratifiées (lorsque K est pair et les paramètres de la j-ème covariable $(\beta^*_{k,j})_{k\in[K]}$ forment deux groupes de taille $K/2$, l'effet « global » n'est pas défini de manière unique, et donc les strates sur lesquelles l'effet diffère de l'effet global non plus). L'approche proposée par [Lee et al., 2013] visant à faire l'inférence conditionnellement au modèle sélectionné pourrait permettre de contourner ce problème.

Enfin, nous envisageons la construction d'un package R implémentant AutoRefLasso sous différents modèles.

4.4.2 AutoRefLasso et modèles de survie à risques compétitifs

Une extension d'AutoRefLasso peut être envisagée pour couvrir les modèles de survie à risques compétitifs, qui apparaissent par exemple naturellement lorsque l'on étudie l'effet de facteurs de risque sur la survenue des différents sous-types de cancer du sein (voir le chapitre introductif de ce document). Dans le cadre de l'étude de risques (ou évènements) compétitifs [Kalbfleisch and Prentice, 2011, Andersen et al., 2012, Aalen et al., 2008], les données proviennent généralement de *cohortes prospectives*. Elles sont utilisées pour décrire l'association entre un vecteur $\mathbf{x} \in \mathbb{R}^p$ de descripteurs (i.e., les facteurs de risque ou encore covariables) et une variable $Y \geq 0$, dite durée de survie, qui mesure le délai entre l'entrée dans l'étude et la survenue d'un évènement d'intérêt. Dans ce type d'étude, la variable Y est le plus souvent censurée à droite : elle n'est pas directement observée et on observe seulement le couple (T, δ). La variable T correspond au temps de suivi, c'est-à-dire le délai entre l'inclusion dans l'étude et le temps de survenue d'un évènement d'intérêt ou d'un évènement dit de censure : $T = \min(Y, C)$, où C est le temps de censure. La variable δ renseigne quant à elle sur le type d'évènement auquel correspond le temps de suivi T : on a ainsi $\delta = 0$ si le temps T correspond à une censure, et $\delta = k$, pour $k \in [K]$ si T correspond à l'évènement d'intérêt k, parmi les $K \geq 2$ évènements d'intérêt considérés dans l'étude [Beyersmann et al., 2011]. Pour chaque individu $i \in [n]$ sain à l'inclusion, nous disposons dans ces études de cohorte des données $(\mathbf{x}_i, T_i, \delta_i)_{i\in[n]}$. Pour simplifier, nous supposerons que la matrice $\mathbf{X}$ renfermant les n observations $\mathbf{x}_i$ est déterministe, que les évènements sont tous indépendants, et que les temps d'évènements sont tous distincts (absence d'ex-aequo).

Une quantité d'intérêt particulier est le risque instantané cause-spécifique, défini pour tout $k \in [K]$ par

$$\lambda_k(t) = \lim_{dt\to 0} \frac{\mathbb{P}(Y \leq t + dt, \delta = k | Y \geq t)}{dt}.$$

Il est classique de considérer la forme suivante pour les risques instantanés cause-spécifiques : pour un individu décrit par les covariables $\mathbf{x} \in \mathbb{R}^p$, on suppose que son risque instantané pour le k-ème évènement au temps t est de la forme [Cox, 1972]

$$\lambda_k(t; \mathbf{x}) = \lambda_{0,k}(t) \exp(\mathbf{x}^T \boldsymbol{\beta}^*_k) \quad \text{pour tout } k \in [K]. \tag{4.13}$$

La fonction $\lambda_{0,k}$ est le risque instantané de base du k-ème évènement. Le terme $\mathbf{x}^T\boldsymbol{\beta}^*_k$ est le prédicteur linéaire pour le k-ème évènement, indépendant du temps, si bien que pour

$\mathbf{x}_1 \neq \mathbf{x}_2$, $\lambda_k(t;\mathbf{x}_1)/\lambda_k(t;\mathbf{x}_2) = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T\boldsymbol{\beta}_k^*)$ est lui-même indépendant du temps. Ce type de modèle appartient ainsi à la famille des modèles à risque proportionnel, tout comme le modèle de Cox [Cox, 1972] qui est couramment utilisé lorsque $K = 1$, c'est-à-dire en présence d'un seul évènement d'intérêt.

Les vecteurs $\boldsymbol{\beta}_k^*$, $k \in [K]$ sont composés des (logarithmes des) *hazard ratio* correspondant à chaque facteur de risque pour le k-ème évènement d'intérêt. Pour les estimer, une approche consiste à utiliser un modèle de Cox sur les données $(T_i, \mathbf{x}_i, \delta_i^{(k)})_{i\in[n]}$ où $\delta_i^{(k)} = \mathbb{I}(\delta_i = k)$ [Beyersmann et al., 2011]. En d'autres termes, on applique le modèle de Cox en considérant que tout évènement autre que le k-ème évènement d'intérêt correspond à une censure. Un estimateur $\hat{\boldsymbol{\beta}}_k$ peut alors être défini comme la solution du problème de maximisation de la *vraisemblance partielle* [Cox, 1972]. Soit $t_1^{(k)} < \ldots < t_{m_k}^{(k)}$ les temps de survenue du k-ème évènement sur notre n-échantillon (on a $0 < m_k \leq n$), et $(i_1^{(k)}), \ldots (i_{m_k}^{(k)})$ les indices de $[n]$ tels que pour tout $\iota \in [m_k]$, $T_{(i_\iota^{(k)})} = t_\iota^{(k)}$ et $\delta_{(i_\iota^{(k)})} = k$. La vraisemblance partielle est alors définie par

$$L_k(\boldsymbol{\beta}_k) = \prod_{\iota\in[m_k]} \frac{\exp(\mathbf{x}_{(\iota)}^T\boldsymbol{\beta}_k)}{\sum_{i\in R_\iota^{(k)}} \exp(\mathbf{x}_i^T\boldsymbol{\beta}_k)},$$

où $R_\iota^{(k)}$ correspond à l'ensemble des individus à risque du k-évènement au temps $t_\iota^{(k)}$, c'est-à-dire l'ensemble des individus pour lesquels $Z_i \geq t_{i_k}$. Pour plus de détails, nous renvoyons le lecteur à [Beyersmann et al., 2011] (chapitre 5), à [Kalbfleisch and Prentice, 2011] (chapitre 8) et à [Lunn and McNeil, 1995]. De manière équivalente, les estimateurs $\hat{\boldsymbol{\beta}}_k$ sont obtenus comme solution maximisant la vraisemble partielle suivante, « combinant » les vraisemblances partielles correspondant à chaque évènement d'intérêt :

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \prod_{k\in[K]} L_k(\boldsymbol{\beta}_k) = \prod_{k\in[K]} \prod_{\iota\in[m_k]} \frac{\exp(\mathbf{x}_{(\iota)}^T\boldsymbol{\beta}_k)}{\sum_{i\in R_\iota^{(k)}} \exp(\mathbf{x}_i^T\boldsymbol{\beta}_k)}.$$

Cette vraisemblance correspond à celle d'un modèle de Cox stratifié sur le vecteur $\mathbf{Z} = (\mathbf{1}_n, 2\cdot\mathbf{1}_n, \ldots, K\cdot\mathbf{1}_n) \in \mathbb{R}^{nK}$ [Therneau and Grambsch, 2000], en considérant les données $(\mathscr{T}, \boldsymbol{\delta}, \boldsymbol{\mathcal{X}})$ définies comme

$$\begin{aligned} \mathscr{T} &= (\mathbf{T}, \ldots, \mathbf{T}) \in \mathbb{R}^{nK} \\ \boldsymbol{\delta} &= (\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(K)}) \in \mathbb{R}^{nK} \\ \boldsymbol{\mathcal{X}} &= \begin{pmatrix} \mathbf{X} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{X} \end{pmatrix} \in \mathbb{R}^{Kn\times Kp}. \end{aligned} \tag{4.14}$$

Remarquons qu'une version simplifiée du modèle de Cox stratifié consiste à supposer que les $\boldsymbol{\beta}_k^*$ sont tous égaux, et que seuls les risques instantanés de base $\lambda_{0,k}$ varient d'une strate à l'autre. Dans ce cas, une estimation du vecteur commun $\breve{\boldsymbol{\beta}}^*$ est obtenue en remplaçant la matrice $\boldsymbol{\mathcal{X}}$ par $\breve{\boldsymbol{\mathcal{X}}} = (\mathbf{X}, \ldots, \mathbf{X})^T \in \mathbb{R}^{nK\times p}$.

Une version pénalisée par la norme L_1 des paramètres de la log-vraisemblance partielle d'un modèle de Cox stratifié peut-être utilisée pour obtenir des estimations creuses de $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$ (en utilisant la matrice $\boldsymbol{\mathcal{X}}$) ou du vecteur commun $\breve{\boldsymbol{\beta}}^*$ (en utilisant la matrice $\breve{\boldsymbol{\mathcal{X}}}$). Le package `penalized` de R permet cette implémentation [Goeman et al., 2012]. En utilisant ce même package, mais avec la matrice $\bar{\boldsymbol{\mathcal{X}}}$ suivante

$$\bar{\boldsymbol{\mathcal{X}}} = \begin{pmatrix} \mathbf{X} & \tau_1^{-1}\mathbf{X} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{X} & \mathbf{0} & \tau_2^{-1}\mathbf{X} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X} & \mathbf{0} & \mathbf{0} & \ldots & \tau_K^{-1}\mathbf{X} \end{pmatrix} \in \mathbb{R}^{Kn \times Kp},$$

on peut implémenter l'extension d'AutoRefLasso qui revient à maximiser le critère suivant

$$\sum_{k \in [K]} \sum_{\iota \in [m_k]} \log \left(\frac{\exp(\mathbf{x}_{(\iota)}^T \boldsymbol{\beta}_k)}{\sum_{i \in R_\iota^{(k)}} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)} \right) - \lambda_1 \left(\|\overline{\boldsymbol{\beta}}\|_1 - \sum_{k \in K} \tau_k \|\boldsymbol{\beta}_k - \overline{\boldsymbol{\beta}}\|_1 \right),$$

sur $(\overline{\boldsymbol{\beta}}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{(K+1)p}$. Comme dans le cas des modèles linéaires (généralisés) sur données stratifiées présenté dans le chapitre précédent, les estimations ainsi obtenues sont typiquement telles que les effets des covariables sur le risque de chaque évènement sont identiques. On a encore bien sûr à l'optimum $\widehat{\overline{\beta}}_j = \text{WSmedian}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j})$.

Ce sujet de l'extension d'AutoRefLasso aux modèles de Cox stratifiés pour traiter le cas des risques compétitifs a donné lieu à un stage de M2 de 4 mois, portant principalement sur l'implémentation de l'approche et une étude de comparaison sur données simulées. Une première application a également été effectuée pour étudier les effets de différents facteurs de risque sur huit sous-types de cancer du sein sur les données de la cohorte E3N. De ces résultats préliminaires, il ressort qu'AutoRefLasso présente les mêmes intérêts que dans le cas des modèles linéaires généralisés, mais que l'implémentation via le package `penalized` se heurte rapidement à des problèmes de mémoire. La suite de ce projet se concentrera dans un premier temps sur la résolution de ces problèmes d'implémentation. Si les problèmes viennent définitivement du package `penalized`, une alternative pourrait être d'utiliser le package `glmnet`. Celui-ci n'implémente pas le modèle de Cox stratifié (et suppose donc que les risques de base $\lambda_{0,k}$ sont tous égaux). On peut cependant l'utiliser sous l'hypothèse que les risques de bases sont proportionnels, i.e., de la forme $\lambda_{0,k}(t) = \alpha_k \lambda_0(t)$. Il suffit alors d'ajouter K colonnes à la matrice $\bar{\boldsymbol{\mathcal{X}}}$, renfermant les K fonctions indicatrices $\mathbb{I}(Z_i = k)$, où Z_i est la i-ème composante du vecteur $\mathbf{Z}$, $i \in [nK]$. D'autre part, un package récent, `penMSM` [Reulen and Kneib, 2015], implémente le fused lasso généralisé dans le contexte des modèles multi-états, dont les modèles de survie à risque compétitifs sont un cas particulier. Ce package devrait ainsi permettre l'implémentation de versions adaptatives d'AutoRefLasso (puisqu'on peut montrer que celles-ci correspondent à des fused lasso généralisés avec un graphe constitué de sous-graphes en étoile, comme dans le cas de RefLasso).

Chapitre 5

Estimation de la structure de modèles graphiques binaires sur données stratifiées

Les modèles de régression (linéaire) considérés jusqu'ici dans ce manuscrit visent à étudier la relation entre une variable réponse, d'intérêt particulier, et des covariables. On est parfois amené à étudier l'ensemble des relations existant au sein d'un groupe de variables, sans se focaliser sur une variable d'intérêt en particulier. C'est notamment le cas dans les exemples cités dans le chapitre introductif de ce manuscrit visant à étudier les associations entre lésions chez les victimes d'accident de la circulation ou entre causes de décès sur les données du CepiDC. Dans ces deux exemples, les variables en jeu sont typiquement binaires : chaque cause est présente ou absente d'un certificat de décès donné, et chaque lésion est de même présente ou absente dans le tableau lésionnel d'une victime d'accident de la circulation. On est alors amené à considérer des modèles graphiques binaires pour représenter la structure de dépendances conditionnelles parmi ces variables. D'autre part, comme nous l'avons décrit dans le chapitre introductif de ce document, ces structures de dépendances peuvent varier en fonction de certaines caractéristiques (âge et sexe dans le cas des certificats de décès, ou encore le type d'usager pour l'étude des lésions) et le problème revient alors à l'estimation simultanée de modèles graphiques binaires sur plusieurs strates prédéfinies de la population.

Dans ce chapitre, nous nous intéressons en premier lieu à l'estimation de la structure d'un seul modèle graphique. Le premier paragraphe présente le modèle d'Ising, qui est classiquement utilisé pour étudier les relations de dépendances conditionnelles parmi un ensemble de variables binaires. Nous présenterons ensuite des approches pénalisées qui permettent de sélectionner les paramètres pertinents de ce modèle. Ces approches, et d'autres qui ne seront pas décrites ici, ont été comparées dans une étude de simulation publiée dans [VV9], où nous proposons également une adaptation d'une des approches qui améliore notablement ses performances. Enfin, le dernier paragraphe présentera les résultats préliminaires de travaux menés pour étendre AutoRefLasso au cas des modèles graphiques binaires et estimer simultanément les modèles correspondant à plusieurs strates de la population. Une application dans le cas de l'étude des associations entre lésions chez les victimes d'accident est proposée pour illustrer cette extension.

5.1 Le modèle d'Ising

Soit $\mathbf{U} = (U_1, ..., U_p)^T \in \{0,1\}^p$ un vecteur p-dimensionnel de variables aléatoires binaires. Etant donné un n-échantillon $\mathbf{U}_1, ..., \mathbf{U}_n$ de répliques i.i.d. de même loi que $\mathbf{U}$, nous souhaitons étudier les associations entre les composantes de $\mathbf{U}$. Une solution réside dans la construction d'un modèle graphique décrivant la loi de probabilité du vecteur $\mathbf{U}$ [Lauritzen, 1996], c'est-à-dire la construction d'un graphe non dirigé $\mathcal{G} = (V, E)$, où V est l'ensemble des p sommets correspondant aux p composantes de $\mathbf{U}$ et l'ensemble d'arêtes $E \subseteq \{(j,\ell) \in V^2 : j < \ell\}$ décrit les relations d'indépendance conditionnelle parmi ces composantes. Plus précisément, l'arête (j,ℓ) entre les variables U_j et U_ℓ de $\mathbf{U}$ est absente si et seulement si U_j et U_ℓ sont indépendantes conditionnellement aux autres variables, contenues dans le vecteur $\mathbf{U}_{-(j,\ell)} \in \mathbb{R}^{p-2}$. La structure du modèle graphique $\mathcal{G}$ correspond à l'ensemble de ses arêtes E. Dans le cadre des modèles graphiques binaires, il est classique de travailler dans la famille des lois de probabilité des modèles exponentiels quadratiques binaires [Cox and Wermuth, 1994, Ravikumar et al., 2010, Banerjee et al., 2008, Höfling and Tibshirani, 2009], ou modèles d'Ising ; notons tout de même que des cas plus généraux peuvent être considérés (voir par exemple [Schwaller et al., 2015]). Sous les modèles d'Ising, on suppose l'existence d'un vecteur de paramètres $\boldsymbol{\theta}^* = ((\theta^*_j)_{1\leq j\leq p}, (\theta^*_{j,\ell})_{1\leq j<\ell\leq p})^T$ de $\mathbb{R}^{p(p+1)/2}$ tel que pour tout vecteur $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$, la probabilité d'observer $\mathbf{U} = \mathbf{u}$ est donnée par

$$\mathbb{P}_{\boldsymbol{\theta}^*}(\mathbf{U} = \mathbf{u}) = \exp\Big\{ \sum_{j=1}^{p} \theta^*_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta^*_{j,\ell} u_j u_\ell - A(\boldsymbol{\theta}^*) \Big\}, \tag{5.1}$$

où la log *partition function* $A : \mathbb{R}^p \to \mathbb{R}$ est définie par

$$A(\boldsymbol{\theta}) = \log \sum_{\mathbf{u}\in\{0,1\}^p} \exp\Big\{ \sum_{j=1}^{p} \theta_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell} u_j u_\ell \Big\}. \tag{5.2}$$

Elle correspond à un terme de normalisation, de telle sorte que $\sum_{\mathbf{u}\in\{0,1\}^p} \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{U} = \mathbf{u}) = 1$ pour tout $\boldsymbol{\theta} \in \mathbb{R}^{p(p+1)/2}$; la convexité stricte de cette fonction assure par ailleurs l'identifiabilité du paramètre $\boldsymbol{\theta}$.

Pour tout $\boldsymbol{\theta} = ((\theta_j)_{1\leq j\leq p}, (\theta_{j,\ell})_{1\leq j<\ell\leq p})^T \in \mathbb{R}^{p(p+1)/2}$, et pour tout $j > \ell$, posons $\theta_{j,\ell} = \theta_{\ell,j}$. Pour tout $j \neq \ell \in [p]^2$, on a sous le modèle (5.1)

$$\frac{\mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 1 | U_\ell = 1, \mathbf{U}_{-(j,\ell)}) / \mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 0 | U_\ell = 1, \mathbf{U}_{-(j,\ell)})}{\mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 1 | U_\ell = 0, \mathbf{U}_{-(j,\ell)}) / \mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 0 | U_\ell = 0, \mathbf{U}_{-(j,\ell)})} = \exp(\theta^*_{j,\ell}). \tag{5.3}$$

Les paramètres $\theta^*_{j,\ell}$ correspondent donc aux log odds-ratios conditionnels et l'indépendance conditionnelle entre les variables U_j et U_ℓ est équivalente à la nullité de $\theta^*_{j,\ell}$. En d'autres termes, l'arête (j,ℓ) est absente du graphe $\mathcal{G}$ si et seulement si $\theta^*_{j,\ell} = 0$. Ainsi, le problème d'estimation de la structure d'un modèle graphique binaire revient, sous le modèle d'Ising, à identifier les paires $(j,\ell) \in [p]^2$, $j < \ell$, pour lesquelles $\theta^*_{j,\ell} = 0$ en (5.1).

On se ramène donc à un problème de sélection de variables dans un modèle paramétrique, qui peut être résolu via des approches pénalisées par la norme L_1 des paramètres par exemple. En notant $\boldsymbol{\mathcal{U}} = (\mathbf{U}_1, ..., \mathbf{U}_n)^T$ la matrice $(n \times p)$ des données, on déduit de (5.1) que la log-vraisemblance pénalisée par la norme L_1 des paramètres s'écrit, pour tout vecteur $\boldsymbol{\theta} \in \mathbb{R}^{p(p+1)/2}$,

$$l(\boldsymbol{\mathcal{U}}; \boldsymbol{\theta}) = \sum_{1 \leq j \leq \ell}^{p} (\boldsymbol{\mathcal{U}}^T \boldsymbol{\mathcal{U}})_{j,\ell} \theta_{j,\ell} - nA(\boldsymbol{\theta}) - n\lambda \|\boldsymbol{\theta}\|_{1,d}, \tag{5.4}$$

où l'on pose $\theta_{j,j} = \theta_j$ et $\|\boldsymbol{\theta}\|_{1,d} = \sum_{j<\ell} |\theta_{j,\ell}|$ (seuls les termes $|\theta_{j,\ell}|$ pour $j < \ell$ sont pénalisés ici puisque la structure du graphe ne dépend pas des termes $\theta_{j,j} = \theta_j$). Cependant, le calcul de la log-vraisemblance (pénalisée ou pas) pour une valeur donnée de $\boldsymbol{\theta}$ requiert celui de la log-partition function $A(\boldsymbol{\theta})$, et donc celui d'une somme sur 2^p termes. Pour des valeurs de $p \geq 20$, ce calcul ne peut pas être effectué en un temps raisonnable et on ne peut donc pas maximiser la vraisemblance (pénalisée ou pas). Diverses solutions approchées ont été proposées dans la littérature. Dans [VV9], nous avons réalisé une revue de la littérature en nous concentrant sur les approches fréquentistes, et principalement sur des approches pénalisées reposant sur une approximation (ou une relaxation) de la vraisemblance des modèles d'Ising. Nous décrivons brièvement certaines de ces méthodes ci-dessous. Nous proposons par ailleurs une modification d'une de ces approches, qui améliore sensiblement ses performances sur l'étude de simulation menée pour comparer ces différentes approches.

5.2 Méthodes approchées pénalisées pour l'estimation de la structure d'un modèle graphique binaire

5.2.1 Régressions logistiques séparées

Une première approche proposée par [Ravikumar et al., 2010] étend celle proposée par [Meinshausen and Bühlmann, 2006] dans le cas des modèles graphiques gaussiens. Elle repose sur l'observation suivante. Pour tout vecteur $\mathbf{u} \in \{0,1\}^p$ et tout $j \in [p]$, soit $\mathbf{u}_{-j} \in \{0,1\}^{p-1}$ le vecteur correspondant au vecteur $\mathbf{u}$ auquel on a ôté la j-ème composante. Sous le modèle (5.1), on a pour tout $j \in [p]$,

$$\text{logit}\{\mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 1 | \mathbf{U}_{-j} = \mathbf{u}_{-j})\} = \theta^*_j + \sum_{\ell \neq j} \theta^*_{j,\ell} u_\ell. \tag{5.5}$$

Pour déterminer quels paramètres $\theta^*_{j,\ell}$ sont nuls dans le modèle (5.1), [Ravikumar et al., 2010] proposent alors d'utiliser p régressions logistiques pénalisées par la norme L_1 de leurs paramètres. Suivant la terminologie introduite par [Wang et al., 2009], nous désignerons cette approche par SepLogit. En se plaçant initialement dans un cadre non-asymptotique, [Ravikumar et al., 2010] établissent des conditions assurant la consistance en sélection de variable de SepLogit. Soit d le degré maximal du graphe, $d = \max_{j \in [p]} |\{\ell \neq j : \theta^*_{j,\ell} \neq 0\}|$. Sous des hypothèses d'incohérence sur la matrice $\boldsymbol{\mathcal{U}}$, ils établissent qu'un nombre d'observations $n > cd^3 \log(p)$, pour une certaine constante $c > 0$, est suffisant pour garantir la consistance en sélection de variables avec grande probabilité. Du point de vue de la théorie

de l'information, cet ordre de grandeur est optimal à un terme d près pour une classe de graphes de degré maximal d [Santhanam and Wainwright, 2012].

Dans SepLogit, p problèmes de régression logistique pénalisés sont résolus séparément. Comme leurs résultats peuvent être asymétriques, au sens où l'on obtient deux estimations pour chaque paramètre $\theta^*_{j,\ell}$, avec en général $\hat{\theta}_{j,\ell} \neq \hat{\theta}_{\ell,j}$, ils doivent être combinés pour estimer la structure de $\mathcal{G}$. Une première possibilité, SepLogit AND, consiste à considérer que l'arête (j,ℓ) est présente dans E si $\hat{\theta}_{j,\ell} \neq 0$ et $\hat{\theta}_{\ell,j} \neq 0$, où $\hat{\theta}_{j,\ell}$ et $\hat{\theta}_{\ell,j}$ sont les estimations de $\theta^*_{j,\ell}$ obtenues en faisant la régression logistique de U_j sur $\mathbf{U}_{-j}$ et de U_ℓ sur $\mathbf{U}_{-\ell}$, respectivement. La deuxième possibilité, SepLogit OR, consiste à considérer que l'arête (j,ℓ) est présente dans E dès lors que $\hat{\theta}_{j,\ell} \neq 0$ ou $\hat{\theta}_{\ell,j} \neq 0$.

On peut contourner ce problème d'asymétrie en ayant recours à la pseudo-vraisemblance [Besag, 1975]. Formellement, la (log-)pseudo-vraisemblance est définie par

$$\sum_{i=1}^{n}\sum_{j=1}^{p} \log\{\mathbb{P}_{\boldsymbol{\theta}}(U_{i,j}|\mathbf{U}_{i,-j})\}, \tag{5.6}$$

pour tout vecteur $\boldsymbol{\theta} \in \mathbb{R}^{p(p+1)/2}$. Ainsi, maximiser la (log-)pseudo-vraisemblance pénalisée par la norme L_1 du vecteur $\boldsymbol{\theta} \in \mathbb{R}^{p(p+1)/2}$ revient à maximiser les p problèmes d'optimisations de SepLogit simultanément sous la contrainte de symétrie $\theta_{j,\ell} = \theta_{\ell,j}$ pour tout $(j,\ell) \in [p]^2$. Un algorithme permettant l'implémentation de cette approche est décrit dans [Höfling and Tibshirani, 2009], et implémenté dans le package `BMN` de R.

5.2.2 Approximation gaussienne de la vraisemblance du modèle d'Ising

Plusieurs approches alternatives reposent sur des « approximations » de la log-partition function [Banerjee et al., 2008, Yang and Ravikumar, 2011]. En particulier, remplaçant la log-partition function par une borne supérieure obtenue par [Wainwright and Jordan, 2008], [Banerjee et al., 2008] dérive un critère approchant le critère (5.4). Il peut de plus être maximisé grâce aux algorithme dédiés à la sélection de covariance, c'est-à-dire à l'identification de la structure d'un modèle graphique gaussien [Dempster, 1972]. Pour tout $i \in [n]$, soit $\mathbf{Z}_i = 2\mathbf{U}_i - 1 \in \{-1,1\}$, $\overline{Z}^{(j)} = (\sum_{i\in[n]} Z_{i,j})/n$ et $\overline{Z} = (\overline{Z}^{(1)},\ldots,\overline{Z}^{(p)})^T \in \mathbb{R}^p$. On définit la matrice de covariance empirique

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i - \overline{Z})(\mathbf{Z}_i - \overline{Z})^T. \tag{5.7}$$

Soit $\lambda \geq 0$ fixé, et soit $\hat{\boldsymbol{\Sigma}}^{-1}_{\lambda}$ la matrice solution du problème d'optimisation suivant

$$\hat{\boldsymbol{\Sigma}}^{-1}_{\lambda} = \underset{\mathbf{M}}{\operatorname{argmax}} \big\{ \log|\mathbf{M}| - \operatorname{tr}(\mathbf{M}(\mathbf{S} + \mathbf{I}_p/3)) - \lambda\|\mathbf{M}\|_1 \big\}, \tag{5.8}$$

où $|\mathbf{M}|$ est le déterminant de la matrice $\mathbf{M}$, $\mathbf{I}_p$ la matrice identité $(p \times p)$, et pour toute matrice symétrique $p \times p$ $\mathbf{M}$, $\|\mathbf{M}\|_1 = \sum_{j<\ell} |M_{j,\ell}|$.

[Banerjee et al., 2008] établissent qu'une solution maximisant leur relaxation de la vraisemblance pénalisée (5.4), pour la valeur λ du paramètre de régularisation, a la forme suivante :

$$\begin{aligned} \hat{\theta}_j &= \overline{Z}^{(j)}, \\ \hat{\theta}_{j,\ell} &= -(\hat{\mathbf{\Sigma}}_\lambda^{-1})_{j,\ell}. \end{aligned} \tag{5.9}$$

Le critère (5.8) correspond à une légère modification du problème de sélection de covariance pénalisé par la norme L_1, où la matrice de covariance empirique est modifiée en ajoutant 1/3 à ses termes diagonaux. Ainsi, tout algorithme dédié au problème de sélection de covariance pénalisé par la norme L_1 peut être utilisé pour estimer la structure d'un modèle graphique binaire. Il suffit de transformer les variables $\{0, 1\}$ en variables $\{-1, 1\}$, ajouter la constante 1/3 aux éléments diagonaux de la matrice de covariance empirique obtenue, et appliquer l'algorithme dédié au cas gaussien, tel que celui implémenté dans le package `glasso` de R par [Friedman et al., 2008]. Nous désignerons cette approche par par GaussCov 1/3 par la suite.

Dans [VV9], nous établissons une connexion entre GaussCov 1/3 et une version de l'approche de [Yang and Ravikumar, 2011] qui repose sur une autre relaxation de la log-partition function. L'approche de [Yang and Ravikumar, 2011] est décrite dans le cadre plus général des variables catégorielles. Dans le cas de variables binaires, et pour certains choix des paramètres intervenant dans cette approche, nous établissons qu'elle revient à identifier la structure du modèle graphique par simple seuillage des covariances empiriques $|S_{j,\ell}|$; elle sera désignée par Cov.Thresh par la suite. L'approche GaussCov 1/3 de [Banerjee et al., 2008] peut-être vue comme le raffinement de l'approche consistant à seuiller les éléments de la matrice de concentration (ou précision), c'est-à-dire l'inverse de la matrice de covariance. En faisant le parallèle avec les distributions gaussiennes multivariées, où les relations d'indépendances conditionnelles parmi les composantes se déduisent des coefficients de corrélation partielles, c'est-à-dire les éléments de la matrice de concentration, travailler avec cette matrice, plutôt que la matrice de covariance, semble mieux adapté lorsque l'on s'intéresse aux relations d'indépendances conditionnelles (et non marginales). Les résultats de notre étude de simulation confirment que l'approche GaussCov 1/3 de [Banerjee et al., 2008] est plus performante que Cov.Thresh. en matière de sélection de la structure des modèles graphiques binaires (sur les configurations considérées dans notre étude de simulation).

Modification de l'approche GaussCov 1/3

Je me suis initialement intéressé aux modèles graphiques binaires pour analyser les associations entre cause de décès sur les certificats de décès à disposition du CepiDC. Face aux nombreuses approches développées dans la littérature, et à l'absence relative d'études comparatives, notamment entre les approches de type SepLogit et celle reposant sur l'approximation gaussienne, nous avons entrepris une étude de simulation. Celle-ci nous a tout d'abord révélé que l'approche GaussCov 1/3 affichait des performances modestes. Dans [VV9], nous avons alors également cherché à l'améliorer, de manière heuristique.

Une première observation est que le terme 1/3 que l'on ajoute à la diagonale de la matrice de covariance empirique est un peu intriguant à première vue. Comme ce terme provient d'une majoration, et non d'une approximation au sens strict, une question naturelle est celle de la performance de l'approche utilisant directement la matrice de covariance S, plutôt que $S + \mathbf{I}_p/3$; nous la désignerons par GaussCov. Nous avons également considéré une autre version, GaussCor, qui utilise la matrice de corrélation plutôt que la matrice de covariance. On peut « justifier » ce choix en remarquant que dans le cas binaire, la statistique du test du chi-deux (pour tester l'indépendance marginale entre deux variables) est $\chi^2 = nr^2$, où r est le coefficient de corrélation de Pearson entre les deux variables binaires considérées (aussi appelé coefficient ϕ) : la corrélation apparaît donc à cet égard comme une meilleure mesure de l'association entre deux variables binaires. D'autre part, les trois versions GaussCov 1/3, GaussCov et GaussCor peuvent se résumer ainsi. On estime le coefficient $\theta^*_{j,\ell}$, $\ell \neq j$, par $-(\widehat{C}^{(\nu)}_\lambda)_{j,\ell}$ où la matrice $\widehat{C}^{(\nu)}_\lambda$ est définie, pour $\nu = 1, 2, 3$, par

$$\widehat{C}^{(\nu)}_\lambda = \arg\max_M \left\{ \log|M| - \mathrm{tr}(MS^{(\nu)}) - \lambda\|M\|_1 \right\} \tag{5.10}$$

avec $S^{(1)} = (S + I_p/3)$, $S^{(2)} = S$ et $S^{(3)} = DSD$, où D est la matrice $p \times p$ diagonale dont le k-ème élément diagonal est $D_{k,k} = 1/\sqrt{S_{k,k}}$.

Or, on peut montrer que travailler dans (5.10) avec la matrice de covariance $S^{(2)} = S$ au lieu de la matrice de corrélation $S^{(3)} = DSD$ revient à travailler avec la matrice de corrélation en remplaçant le terme de pénalité $\sum_{k<\ell} |M_{k,\ell}|$ par $\sum_{k<\ell} |M_{k,\ell}/(\sqrt{S_{kk}S_{\ell\ell}})|$. Autrement dit, les associations entre variables dont le produit des variances est faible (et donc dont les prévalences sont soit élevées soit faibles) sont plus fortement pénalisées lorsqu'on utilise la matrice de covariance S en (5.10). Au vu du lien entre le coefficient de corrélation et la statistique du χ^2, la corrélation dépend à la fois de la force de l'association entre deux variables (mesurées par l'odds-ratio par exemple) et du produit de leurs variances. Il ne semble donc pas nécessaire de pénaliser plus fortement les associations entre variables de faibles variances, et GaussCor nous est apparu pertinent à cet égard.

5.2.3 Comparaison sur données simulées

Dans [VV9], nous avons comparé, sur données simulées, les approches décrites ci-dessus, ainsi que celle reposant sur le seuillage de l'information mutuelle conditionnelle (CMIT, pour Conditional Mutual Information Thresholding) décrite dans [Anandkumar et al., 2012]. L'objectif premier de ce travail était l'application sur les données du CepiDC pour étudier les associations entre les causes de décès (selon une catégorisation à 59 causes) dans les certificats de décès, sur différents sous-groupes définis par les classes d'âge et le sexe de la personne décédée. Dans cette application, le nombre d'observations est au minimum de l'ordre du millier, avec $p = 60$. Nous avons donc cherché à évaluer les différentes approches dans ce cadre où n est grand devant p.

Nous avons considéré différentes configurations dans [VV9] et la table 5.1 présente certains des résultats obtenus pour $p = 10$ et $p = 50$, et différentes valeurs de n. Pour chaque approche et chaque jeu de données simulé, nous calculons le temps nécessaire à la résolution numérique, la précision de l'identification du support (Acc.) et le F1-score. La sélection des

paramètres de régularisation a été opérée via un critère de type 2StepBIC. Les résultats de la table 5.1 correspondent aux moyennes de ces critères sur 50 réplications.

Premièrement, dans le cas $p = 10$, la comparaison entre GaussCor et GaussCov 1/3 illustre bien les problèmes de GaussCov 1/3, qui n'est pas assez sensible : GaussCov 1/3 détecte moins d'associations que GaussCor et affiche des valeurs modestes pour le F1-score notamment. Ce défaut est partagé par l'approche Cov.Thresh. et pourrait donc être imputable à l'utilisation des covariances plutôt que les corrélations. Les autres approches fournissent des modèles aux performances comparables. En particulier, GaussCor atteint des performances au moins comparables aux autres approches, et corrige donc les défauts de GaussCov 1/3. Les différences les plus notables entre ces méthodes concernent les temps de calcul. En particulier, GaussCor est très rapide dans les cas présentés ici. Lorsque $p = 200$ cependant, nous obtenons dans [VV9] des résultats qui viennent tempérer cette observation : les approches SepLogit sont alors plus rapides que GaussCor. Deux remarques peuvent compléter ces comparaisons sur les temps de calcul de SepLogit et GaussCor. D'une part, SepLogit est implémentée en utilisant le package `glmnet`, qui incorpore une étape d'élimination de features a priori (selon une méthode voisine de l'approche SaFe présentée au chapitre 2). La fonction glasso utilisée pour l'implémentation de GaussCor n'incorpore pas encore cette option : la comparaison des temps de calcul est en ce sens à l'avantage de SepLogit. D'autre part, on peut facilement paralléliser SepLogit (puisqu'elle repose sur la résolution de p régressions logistiques pénalisées indépendantes), ce qui n'a pas été fait ici et les temps de calcul peuvent donc facilement être divisés par $\min(p, Q)$ pour SepLogit, où Q désigne le nombre de coeurs disponibles sur la machine. Notons enfin que l'approche reposant sur la pseudo-vraisemblance affichait des performances analogues à SepLogit en matière de sélection de variables, mais des temps de calcul beaucoup plus longs. Sa parallélisation est d'autre part moins directe que pour SepLogit, les p vraisemblances étant maximisées conjointement sous la contrainte de symétrie $\theta_{j,\ell} = \theta_{\ell,j}$.

Une dernière remarque concerne la cohérence des associations détectées par les différentes approches. Sur les configurations considérées dans notre étude de simulation, SepLogit et GaussCor, par exemple, renvoient des modèles aux performances comparables. Cependant, nous avons observé que les associations détectées par chacune de ces approches pouvaient différer sensiblement sur un même jeu de données. Dans de tels cas, une solution peut consister à retourner l'intersection des associations détectées par SepLogit OR et GaussCor, ou l'union des associations détectées par SepLogit AND et GaussCor par exemple. Nous avons évalué ces deux stratégies dans [VV9] : ces deux stratégies affichent des performances comparables à GaussCor et SepLogit, tout en limitant les taux de faux positifs (lorsqu'on prend l'intersection) ou de faux négatifs (lorsqu'on prend l'union).

5.3 Estimation simultanée de la structure de plusieurs modèles graphiques binaires

Un de mes projets en cours concerne l'estimation conjointe de plusieurs modèles graphiques binaires. Ce projet est né de l'analyse des associations entre causes de décès sur les données du CepiDC, mais une autre application intéressante concerne la description des

TABLE 5.1 – Résultats de la comparaison empirique des méthodes. Les moyennes (et écart-type), calculés à partir de 50 réplications, sont donnés pour les temps de calcul en secondes, le nombre d'associations détectées, la précision quant à l'identification du support et le F1-score correspondant.

(a) $p = 10$

	$n = 100$			$n = 2500$		
Method	Comp. Time	Acc.	F1 score	Time (s)	Acc.	F1-score
Cov.Thresh.	5.70 (0.07)	0.77 (0.07)	0.05 (0.09)	117.84 (1.12)	0.81 (0.07)	0.30 (0.16)
SepLogit AND	1.00 (0.17)	0.77 (0.07)	0.10 (0.11)	5.19 (0.43)	0.86 (0.06)	0.57 (0.15)
SepLogit OR	1.00 (0.17)	0.77 (0.06)	0.17 (0.12)	5.19 (0.43)	0.87 (0.05)	0.62 (0.15)
GaussCor	0.06 (0.01)	0.77 (0.06)	0.13 (0.13)	0.06 (0.01)	0.87 (0.05)	0.62 (0.13)
GaussCov 1/3	0.06 (0.01)	0.77 (0.07)	0.04 (0.07)	0.06 (0.01)	0.82 (0.07)	0.35 (0.18)
CMIT	0.28 (0.05)	0.77 (0.07)	0.14 (0.14)	0.32 (0.04)	0.87 (0.05)	0.61 (0.13)

(b) $p = 50$

	$n = 500$			$n = 2500$		
Method	Comp. Time	Acc.	F1 score	Time (s)	Acc.	F1-score
SepLogit AND	14.07 (1.39)	0.96 (0.01)	0.15 (0.07)	23.16 (0.69)	0.97 (0.01)	0.50 (0.12)
SepLogit OR	14.07 (1.39)	0.95 (0.01)	0.19 (0.08)	23.16 (0.69)	0.97 (0.01)	0.55 (0.10)
GaussCor	0.79 (1.61)	0.95 (0.01)	0.18 (0.07)	1.07 (2.26)	0.97 (0.01)	0.56 (0.10)
CMIT	34.81 (6.82)	0.95 (0.01)	0.17 (0.08)	46.02 (5.66)	0.97 (0.01)	0.55 (0.10)

associations entre les lésions subies chez les victimes d'accident de la circulation, en fonction des caractéristiques de l'accident. Dans un premier temps, et pour illustrer le propos, nous nous concentrerons sur les caractéristiques décrivant simplement le type d'usager touché. En d'autres termes, la question est de déterminer les profils d'associations entre lésions chez les victimes d'accident de la circulation en fonction du type d'usager, et notamment déterminer si ces profils d'associations varient en fonction du type d'usager.

L'estimation conjointe de $K \geq 1$ modèles graphiques revient à estimer l'ensemble des $Kp(p+1)/2$ paramètres $\theta^{(k)*}_{j_1,j_2}$, pour $k \in [K]$ et $(j_1, j_2) \in [p]^2$ avec $j_1 \leq j_2$. Comme dans le cas de l'estimation de modèles de régression sur données stratifiées, la plupart des applications concernent des modèles graphiques pour lesquels la structure varie peu avec $k \in [K]$. Dans le cas gaussien, [Danaher et al., 2014] propose alors une pénalité de type fused lasso généralisé pour encourager les modèles à partager la même structure (une pénalité de type group lasso est également proposée). Pour les modèles graphiques binaires, une approche analogue est proposée par [Ahmed and Xing, 2009] pour estimer les structures sur des périodes de temps successives : les auteurs utilisent l'approche SepLogit avec une pénalité L_1 et une pénalité fused pour encourager les similarités entre les modèles correspondant à des années consécutives. Récemment, [Guo et al., 2015] ont proposé une approche alternative reposant sur l'utilisation de la pseudo-vraisemblance avec une décomposition multiplicative des paramètres $\theta^{(k)}_{j_1,j_2} = \bar{\theta}_{j_1,j_2}\gamma^{(k)}_{j_1,j_2}$ et une pénalisation des termes $|\bar{\theta}_{j_1,j_2}|$ et $|\gamma^{(k)}_{j_1,j_2}|$. Comme dans la décomposition additive que nous utilisons dans AutoRefLasso, le terme $\bar{\theta}_{j_1,j_2}$ peut être vu comme le niveau d'association global entre les variables j_1 et j_2,

et $\gamma^{(k)}_{j_1,j_2}$ mesure la différence entre ce niveau global et le niveau d'association dans la k-ème strate. Cette décomposition multiplicative a été proposée par [Lozano and Swirszcz, 2012] dans le modèle linéaire. Combinée aux pénalisations des termes $|\bar{\theta}_{j_1,j_2}|$ et $|\gamma^{(k)}_{j_1,j_2}|$ elle encourage les associations à être nulles sur l'ensemble des strates (si $\hat{\bar{\theta}}_{j_1,j_2} = 0$) ou sur certaines strates seulement (si $\hat{\gamma}^{(k)}_{j_1,j_2} = 0$). Par contre, si l'approche retourne des estimations non nulles pour les niveaux d'association entre les variables j_1 et j_2 sur les strates k_1 et k_2, $\hat{\theta}^{(k_1)}_{j_1,j_2} \neq 0$ et $\hat{\theta}^{(k_2)}_{j_1,j_2} \neq 0$, alors on a $\hat{\theta}^{(k_1)}_{j_1,j_2} \neq \hat{\theta}^{(k_2)}_{j_1,j_2}$ par construction et cette approche ne semble donc que modérément adaptée lorsque la question principale est la détection des hétérogénéités. L'utilisation de SepLogit, par exemple, avec une pénalité de type fused lasso généralisé ou celle utilisée dans AutoRefLasso, semble mieux adaptée. Un stagiaire de M1, Alexei Novoloaca (Master Santé Publique de l'Université Lyon 1, option biostatistique) a déjà travaillé avec moi sur l'implémentation de l'approche de [Guo et al., 2015] et l'extension d'AutoRefLasso pour l'estimation conjointe de plusieurs modèles graphiques binaires. Des résultats de simulation préliminaires soulignent la bonne tenue d'AutoRefLasso dans ce contexte.

Par ailleurs, un autre stagiaire de M1, Yacine Berkane (Polytech. Lyon), avait quant à lui travaillé sur une représentation graphique adaptée pour comparer visuellement les structures de plusieurs modèles graphiques. Afin de faciliter ces comparaisons, nous avons opté pour une représentation où la position de chacun des noeuds du graphe (les lésions dans notre exemple) est commune sur chaque strate. D'autre part, un code couleur permet de distinguer les lésions en fonction de leur zone corporelle (tête et cou, membres supérieurs, colonne, thorax, abdomen, membres inférieurs, etc.). Chaque lésion est représenté par un disque, dont la surface est proportionnelle à sa fréquence sur la strate considérée. Enfin, les associations sont représentées par des arêtes dont l'épaisseur est proportionnelle au niveau d'association (mesuré par l'odds-ratio conditionnel $\exp(\hat{\theta}^{(k)}_{j,\ell})$. En utilisant le code R développé par Yacine, Alexei a fait une première application d'AutoRefLasso sur les données du Registre du Rhône pour illustrer l'approche ; ces données décrivent notamment l'ensemble des lésions subies par les victimes d'accident de la circulation survenues dans le Rhône entre 1996 et 2013. La figure 5.1 présente ces résultats préliminaires et décrit les associations entre lésions chez quatre types d'usagers : les automobilistes, les usagers de de deux-roues motorisés (2RM), les piétons et les cyclistes. A noter que nous ne représentons ici que les associations retournées positives, avec $\hat{\theta}^{(k)}_{j,\ell} > \log(1.5)$. A la lecture de ces graphes, plusieurs résultats sont marquants. Par exemple, les lésions à la tête (en gris) sont moins fréquentes chez les usagers de 2RM que chez les autres usagers (les surfaces des disques sont plus faibles), ce qui peut s'expliquer par la protection de la tête induite par le port du casque. Cependant, les profils d'associations entre lésions à la tête sont relativement similaires d'un type d'usager à l'autre. Cela souligne que le casque protège effectivement les lésions au crâne chez les usagers de 2RM, mais qu'à partir du moment où une lésion à la tête survient quand même chez un usager de ce type (soit parce qu'il ne porte pas de casque, soit parce que le niveau de protection du casque était trop faible par rapport à la violence du choc subi), le tableau des lésions touchant la tête est analogue à celui des autres usagers ; on observe même des associations légèrement plus fortes. Un autre résultat intéressant est

que le graphe décrivant les automobilistes est le plus dense. En particulier, on détecte chez les automobilistes des associations entre lésions des membres inférieurs, entre lésions des membres supérieurs, entre lésions de ces deux zones corporelles, et également entre lésions des membres inférieurs et du thorax, que l'on détecte beaucoup moins (voire pas du tout) chez les autres usagers. Ces associations illustrent ce que les traumatologues appellent le « syndrome du tableau de bord » : les conducteurs qui percutent violemment le tableau de bord subissent typiquement des lésions multiples aux membres inférieurs, aux membres supérieurs et au thorax.

Ces résultats préliminaires ont beaucoup intéressé les traumatologues de l'UMRESTTE. Pour qu'ils aient un réel intérêt clinique, il est cependant nécessaire d'aller plus loin, selon différentes directions. Nous avons ici utilisé une catégorisation des lésions en 27 classes, partant d'une catégorisation à plus de 1300 classes et il est donc incontournable de réfléchir à une catégorisation plus fine. D'autre part, on pourrait également affiner la définition des strates en y incorporant le type d'antagoniste (pour les accidents à plusieurs véhicules), la gravité de l'accident, etc. Toutes ces pistes seront creusées dans les mois à venir.

D'un point de vue méthodologique, on pourrait également chercher à étendre AutoRefLasso autrement qu'en le combinant à SepLogit. Considérant dans un premier temps le cas des modèles graphiques gaussiens, on pourrait chercher à décomposer les matrices de précisions $\mathbf{\Theta}^{(k)}$ décrivant les associations sur chaque strate en la somme suivante :

$$\mathbf{\Theta}^{(k)} = \bar{\mathbf{\Theta}} + \mathbf{\Gamma}^{(k)},$$

et maximiser, en $\bar{\mathbf{\Theta}}$ et $(\mathbf{\Gamma}^{(k)})_{k\in[K]}$, le critère suivant :

$$\sum_{k\in[K]} \big\{ \log|\bar{\mathbf{\Theta}} + \mathbf{\Gamma}^{(k)}| - \mathrm{tr}((\bar{\mathbf{\Theta}} + \mathbf{\Gamma}^{(k)})\mathbf{S}^{(k)}) - \lambda_1\|\bar{\mathbf{\Theta}}\|_1 - \lambda_{2,k}\|\mathbf{\Gamma}^{(k)}\|_1 \big\},$$

avec $\mathbf{S}^{(k)}$ la matrice de covariance empirique de la k-ème strate.

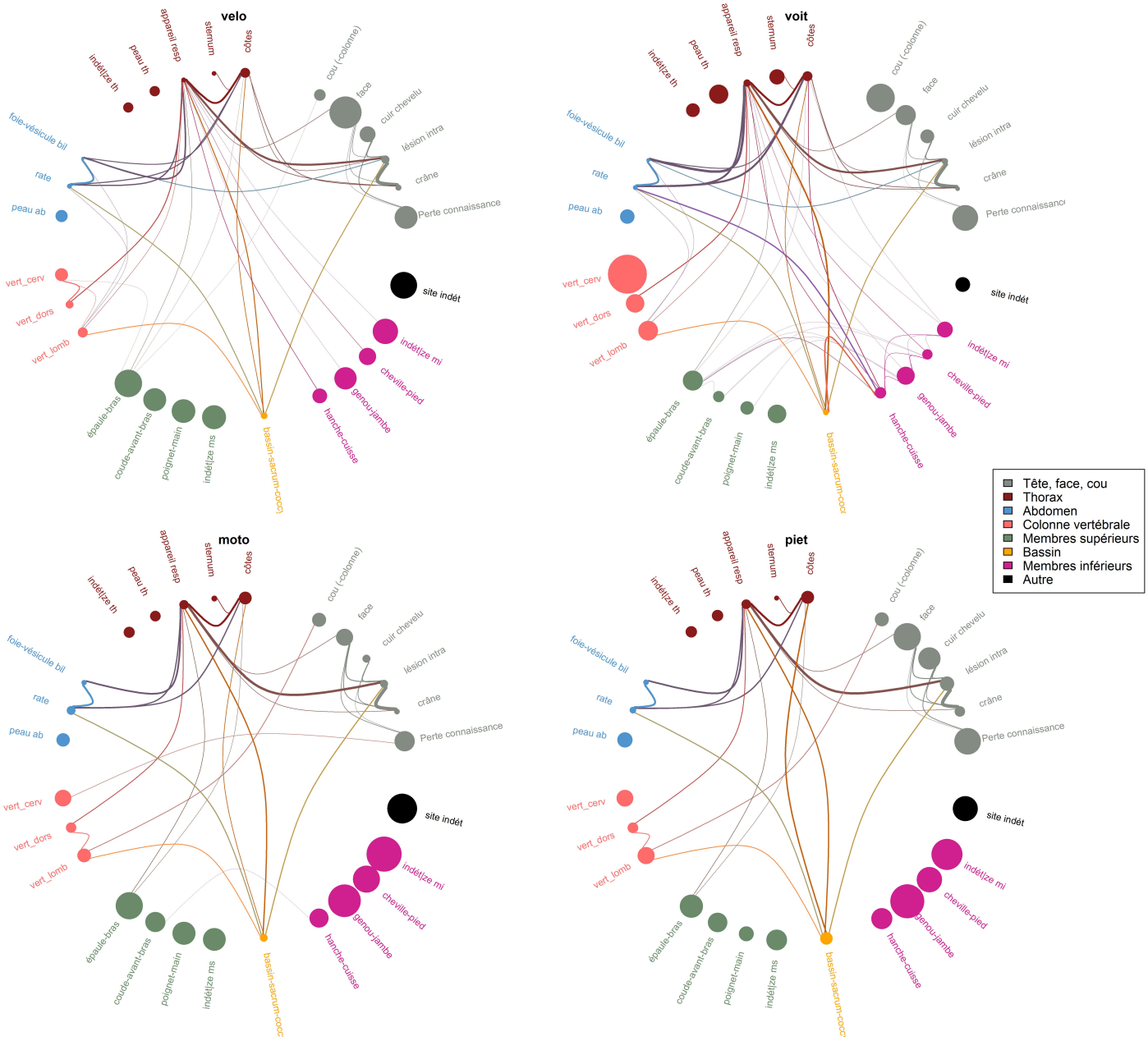


FIGURE 5.1 – Application d'AutoRefLasso pour estimer la structure de plusieurs modèles graphiques binaires. Résultats préliminaires sur les données du Registre du Rhône pour étudier les profils d'associations entre lésions chez les victimes d'accident de la circulation, en fonction du type d'usager : cycliste (velo), automobiliste (voit), motard (moto) et piéton (piet).

Troisième partie

Causalité sur données observationnelles

Chapitre 6

Causalité et responsabilité en sécurité routière

6.1 Introduction

Ce chapitre décrit un projet en cours qui fait l'objet de la thèse de Marine Dufournet, que je co-encadre avec Jean-Louis Martin (CR, IFSTTAR) et Alain Bergeret (PUPH, UCBL). En matière de sécurité routière, la plupart des causes d'accident en lien avec les usagers de la route sont considérées comme établies : alcoolémie, vitesse, usage du téléphone au volant, drogue, médicament, etc. La question posée aux épidémiologistes est donc maintenant celle de la quantification des effets de ces causes, en particulier sur le risque d'accident.

Les outils développés dans la littérature en lien avec l'inférence causale permettent de déterminer précisément les conditions sous lesquelles les effets causaux d'une cause connue peuvent être identifiés et estimés à partir des données disponibles (voir le paragraphe 6.2), voire décomposés en effets direct et indirect en présence de médiateurs (voir paragraphe 6.3). Cependant, une difficulté particulière lorsque l'on s'intéresse aux causes des accidents provient du fait que les données disponibles ne concernent généralement que des usagers impliqués dans des accidents : l'absence de données relatives aux témoins (les usagers circulants) rend impossible l'estimation des effets sur le risque d'accident. Même si d'autres types d'analyse ont été proposés, il est maintenant classique d'effectuer des *analyses en responsabilité* [Brubacher et al., 2014, Salmi et al., 2014]. Celles-ci reposent sur la connaissance du niveau de responsabilité de chacun des conducteurs impliqués dans l'accident. Le paragraphe 6.4 présente l'état de nos réflexions quant à l'identification des effets causaux dans les analyses en responsabilité.

6.2 Effet causal et variables contrefactuelles

Pour simplifier l'exposé, nous nous focalisons ici sur le cas de deux variables X et Y binaires à valeurs dans $\{0, 1\}$. A ce jour, diverses conceptions de la causalité co-existent [Chambaz et al., 2014]. En premier lieu, la conception régulariste de Hume [Hume, 1739] considère que la cause est toujours suivie de son effet. Cette conception a ensuite été étendue en considérant qu'une cause est une condition INUS, acronyme de l'anglais Insufficient but Nonredundant part of an Unnecessary but Sufficient (condition) [Mill, 1856, Mackie, 1974].

Depuis la fin du 20ème siècle, la conception probabiliste de la causalité incorpore la notion de hasard : l'évènement $\{X = 1\}$ est une cause de $\{Y = 1\}$ si et seulement si $\{X = 1\}$ augmente la probabilité de $\{Y = 1\}$, *toutes choses égales par ailleurs* (voir [Chambaz et al., 2014, Greenland et al., 1999, Pearl, 2009, Robins, 1986, Rubin, 1974, Rothman et al., 2008]). Le « toutes choses égales par ailleurs » se rapporte ici aux lois probabilistes du modèle causal conduisant potentiellement à l'évènement $\{Y = 1\}$ et non à un simple conditionnement, par exemple sur un évènement défini à partir d'un ensemble de facteurs de confusion. Le modèle causal peut être décrit par un graphe orienté, qu'on supposera acyclique, et qu'on notera DAG (Directed Acyclic Graph). Trois exemples simples sont donnés en figure 6.1. Dans chacun des cas, on peut décrire le système $O = (W, X, Y)$ par trois équations structurelles, à l'aide de trois fonctions déterministes f_W, f_X, f_Y et trois variables aléatoires indépendantes U_W, U_X, U_Y, ou perturbations (voir la légende de la figure 6.1). Ainsi combinés, le DAG et ces équations structurelles forment un modèle causal structurel (Structured Causal Model, SCM). Ces modèles ont été largement développés par Pearl ([Pearl, 2000, Pearl, 2009]).

Dans les SCMs, on peut associer au système « naturel » O son pendant « contrôlé » $O(x)$ que l'on aurait observé, dans un monde possiblement contrefactuel, si l'on avait imposé la valeur x à X. Pour décrire plus précisément ce système $O(x)$, nous avons recours aux variables contrefactuelles, ou résultats potentiels. En particulier, on peut définir les variables $Y(0)$ et $Y(1)$ issues du système $O(0)$ et $O(1)$ respectivement, que l'on aurait observé si l'on avait imposé $X = 0$ et $X = 1$ respectivement. Dans le cadre des SCMs, la variable $Y(x)$ est définie précisément, via la même fonction déterministe f_Y que la variable Y, mais en modifiant certains arguments de cette fonction : X devient x, etc. Quelques exemples simples sont donnés dans le paragraphe suivant. En particulier, l'hypothèse dite de cohérence, selon laquelle $Y = XY(1) + (1 - X)Y(0)$ ou encore $Y = Y(X)$, est directement vérifiée sous les SCMs. Elle s'interprète comme « la coïncidence de l'issue dans le monde actuel avec l'issue dans le monde contrefactuel exploré » [Chambaz et al., 2014]. Sous cette hypothèse, l'inférence causale peut être vue comme un problème de données manquantes : l'effet causal de X sur Y se définit à partir des variables $Y(0)$ et $Y(1)$, qui ne sont que partiellement observées. On peut par exemple considérer l'excès de risque « moyen »

$$\mathbb{E}(Y(1) - Y(0)) = \mathbb{P}(Y(1) = 1) - \mathbb{P}(Y(0) = 1).$$

En général, on a $\mathbb{P}(Y(x) = 1) \neq \mathbb{P}(Y = 1|X = x)$ pour $x \in \{0, 1\}$, et l'enjeu de l'inférence causale sur données observationnelles est de décrire les situations où les quantités $\mathbb{P}(Y(x) = 1)$ sont identifiables [1] [Pearl, 2000].

Sous l'hypothèse dite d'ignorabilité, à savoir $(Y(0), Y(1)) \perp\!\!\!\perp X$, on a

$$\mathbb{E}(Y(x)) = \mathbb{P}(Y(x) = 1) = \mathbb{P}(Y(x) = 1|X = x) = \mathbb{P}(Y = 1|X = x).$$

En d'autres termes, un échantillon représentatif des évènements $\{X = 0\}$ et $\{X = 1\}$ suffit pour estimer sans biais l'effet causal de X sur Y, sous l'hypothèse d'ignorabilité. Cette hypothèse est en particulier vérifiée sous l'hypothèse de randomisation, et donc dans l'essai

1. $\mathbb{P}(Y(x) = y)$ est dite identifiable si les hypothèses induites par la structure du DAG G assurent que cette quantité peut être exprimée à partir de la distribution des variables observées $\boldsymbol{\mathcal{V}}$ qui composent G ; voir la définition 1 de [Bareinboim and Tian, 2015] par exemple.

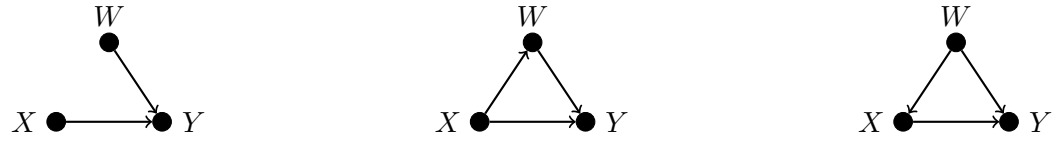


FIGURE 6.1 – Trois exemples de DAG décrivant le système conduisant potentiellement à l'évènement $\{Y = 1\}$. La cause potentielle est notée X. La variable W représente quant à elle une troisième variable dont le rôle dans le système dépend du DAG. Dans chacun des trois DAGs, Y est causée par W et X. Dans le DAG de gauche, il n'existe aucune relation causale entre X et W et les 3 équations structurelles sont $W = f_W(U_W)$, $X = f_X(U_X)$ et $Y = f_Y(X, W, U_Y)$. Dans le DAG du milieu, X est une cause de W : W est considéré comme un facteur intermédiaire et les 3 équations structurelles sont $X = f_X(U_X)$, $W = f_W(X, U_W)$ et $Y = f_Y(X, W, U_Y)$. Dans le DAG de droite, W est une cause de X : W est alors considéré comme un facteur de confusion et les 3 équations structurelles sont $W = f_W(U_W)$, $X = f_X(W, U_X)$ et $Y = f_Y(X, W, U_Y)$.

thérapeutique où l'expérimentateur intervient directement sur la variable X, de manière aléatoire. La variable X est alors indépendante de toute autre variable potentiellement liée à Y, comme dans le DAG de gauche de la figure 6.1. Dans ce cas, le système contrôlé $O(x)$ est décrit par les équations structurelles : $X = x$, $W = f_W(U_W)$ et $Y(x) = f_Y(x, W, U_Y)$. Comme W (et U_Y) sont indépendants de X, on a bien $(Y(0), Y(1)) \perp\!\!\!\perp X$. L'hypothèse d'ignorabilité est également vérifiée dans le second DAG, puisque les équations structurelles décrivant le système contrôlé $O(x)$ sont : $X = x$, $W(x) = f_W(x, U_W)$ et $Y(x) = f_Y(x, W(x), U_Y)$. Comme dans le cas précédent, on peut montrer que $W(x) \perp\!\!\!\perp X$, et par suite que $(Y(0), Y(1)) \perp\!\!\!\perp X$. Par contre, la condition d'ignorabilité n'est pas vérifiée dans le DAG de droite qui décrit le cas de l'existence d'un facteur de confusion. En effet, les équations structurelles décrivant le système contrôlé $O(x)$ sont : $W = f_W(U_W)$, $X = x$, et $Y(x) = f_Y(x, W, U_Y)$. Cette fois, $Y(x)$ et X ne sont pas indépendants car W et X ne le sont pas.

Cependant, l'hypothèse d'ignorabilité conditionnelle $(Y(0), Y(1)) \perp\!\!\!\perp X|W$ est vérifiée dans ce cas, si bien que

$$\begin{aligned}
\mathbb{E}(Y(x)) &= \mathbb{P}(Y(x) = 1) \\
&= \mathbb{E}_W\big[\mathbb{P}(Y(x) = 1|W\big] \\
&= \mathbb{E}_W\big[\mathbb{P}(Y(x) = 1|W, X = x)\big] \\
&= \mathbb{E}_W\big[\mathbb{P}(Y = 1|W, X = x)\big].
\end{aligned}$$

Sous cette hypothèse d'ignorabilité conditionnelle, et si de plus $0 < \mathbb{P}(X = x|W) < 1$, un échantillon représentatif de la population permet donc d'estimer l'effet causal. Afin d'illustrer la différence entre l'effet causal et les effets estimés dans les analyses d'association classiques, considérons l'exemple simple du modèle linéaire en présence d'un facteur de confusion et sous l'hypothèse d'ignorabilité conditionnelle. Supposons alors que $\mathbb{P}(Y = 1|W, X) = \alpha + \beta_1 X + \beta_2 W + \gamma XW$, pour des paramètres α, β_1, β_2 et γ réels. Si $\gamma = 0$, alors la

formule précédente indique que l'excès de risque causal de X sur Y vaut $\mathbb{E}(Y(1))-\mathbb{E}(Y(0)) = \beta_1$. On retombe donc sur le paramètre associé à la variable X dans le modèle multivarié, ajusté sur W. Cependant, si $\gamma \neq 0$, alors $\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) = \beta_1 + \gamma\mathbb{E}W$ et l'effet causal de X correspond à l'effet de X « moyenné » sur l'ensemble de la population.

L'hypothèse d'ignorabilité conditionnelle est évidemment très forte, et rarement vérifiée en pratique puisqu'elle implique que l'ensemble des facteurs de confusion entre X et Y sont connus et observés. Pour autant, elle ne doit pas être considérée comme une limite spécifique à l'inférence causale. L'inférence causale a surtout pour objectif d'établir les conditions sous lesquelles les effets causaux peuvent être déduits des mesures d'association : si des facteurs de confusion ne sont pas observés, les mesures d'association ajustées sur les facteurs observés n'ont pas d'interprétation causale. D'autre part, l'ajustement sur les facteurs de confusion n'est qu'une des approches possibles pour exprimer les effets causaux à partir de variables observées, et certaines techniques permettent d'estimer des effets causaux en situation de facteurs de confusions non observés : on peut citer par exemple le critère front-door de [Pearl, 1995] (voir aussi [Tian and Pearl, 2002] et [Pearl, 2009]).

Dans le paragraphe suivant, nous présentons brièvement un autre intérêt de l'inférence causale et du recours aux variables contrefactuelles. Elles permettent de déterminer une décomposition de l'effet causal d'un facteur en une somme de deux termes en présence d'un médiateur : effet direct et effet indirect. Notons que les variables contrefactuelles permettent également de définir précisément d'autres mesures classiques en épidémiologie telles que la fraction attribuable [Pearl, 2000].

6.3 Décomposition de l'effet total en présence d'un médiateur

Considérons pour simplifier la situation décrite dans le deuxième DAG de la figure 6.1, et notons alors M la variable qui y était notée W. Dans ce type de modèle causal, cette variable est classiquement appelée un médiateur. Sur l'échelle de l'excès de risque, l'effet causal de X sur Y est défini par $\mathbb{P}(Y(1) = 1) - \mathbb{P}(Y(0) = 1)$. Nous le noterons ATE, pour Average Total Effect. Nous pouvons décomposer cet effet en une somme de deux termes, l'effet direct (NDE, Natural Direct Effet) et l'effet indirect (NIE, Natural Indirect Effect). Notons $M(x)$ la variable aléatoire correspondant au médiateur M que l'on aurait observée dans le monde contrefactuel où l'on aurait imposé $X = x$. Pour tout $(x_1, x_2) \in \{0,1\}^2$, on note enfin $Y(x_1, M(x_2)) = f_Y(x_1, M(x_2), U_Y)$ la variable correspondant à Y que l'on aurait observée après avoir fixé X à la valeur x_1 et M à $f_M(x_2, U_M)$. On a alors $Y(x) = Y(x, M(x))$ si bien que

$$\begin{aligned}
\text{ATE} &= \mathbb{P}(Y(1) = 1) - \mathbb{P}(Y(0) = 1) \\
&= \mathbb{E}[Y(1, M(1)) - Y(0, M(0))] \\
&= \mathbb{E}[Y(1, M(1)) - Y(1, M(0)) + Y(1, M(0)) - Y(0, M(0))] \\
&= \text{NIE}(1) + \text{NDE}(0) \qquad (6.1) \\
&= \mathbb{E}[Y(1, M(1)) - Y(0, M(1)) + Y(0, M(1)) - Y(0, M(0))] \\
&= \text{NDE}(1) + \text{NIE}(0) \qquad (6.2)
\end{aligned}$$

avec $\text{NDE}(x) = \mathbb{E}[Y(1, M(x)) - Y(0, M(x))]$ et $\text{NIE}(x) = \mathbb{E}[Y(x, M(1)) - Y(x, M(0))]$. La quantité $\text{NDE}(x)$ mesure l'augmentation du risque moyen lorsque le médiateur est maintenu à la valeur $M(x)$, alors qu'on force la variable X à passer de 0 à 1 : il s'agit donc bien d'une mesure (ou plutôt deux mesures puisque $x \in \{0, 1\}$) de l'effet direct. De même les quantités $\text{NIE}(x)$ pour $x \in \{0, 1\}$ représentent deux mesures de l'effet indirect de X, médié par M. Sous l'hypothèse d'ignorabilité séquentielle (qui généralise l'hypothèse d'ignorabilité conditionnelle), ces quantités sont identifiables à partir des variables observées X, M et Y [Imai et al., 2010]. Plus précisément, on obtient, sous cette hypothèse (et en l'absence de facteurs de confusion),

$$\begin{aligned}
\text{NDE}(x) &= \sum_{m\in\{0,1\}} \big[\mathbb{P}(Y = 1|X = 1, M = m) - \mathbb{P}(Y = 1|X = 0, M = m)\big]\mathbb{P}(M = m|X = x) \\
\text{NIE}(x) &= \sum_{m\in\{0,1\}} \mathbb{P}(Y = 1|X = x, M = m)\big[\mathbb{P}(M = m|X = 1) - \mathbb{P}(M = m|X = 0)\big].
\end{aligned}$$

En d'autres termes, il suffit d'estimer les probabilités conditionnelles du type $\mathbb{P}(Y = 1|X, M)$ et $\mathbb{P}(M = 1|X)$ pour estimer les quantités $\text{NDE}(x)$ et $\text{NIE}(x)$.

6.4 Effets causaux dans les analyses en responsabilité

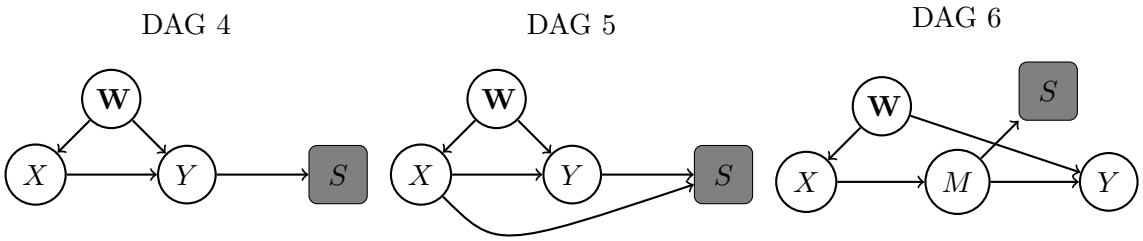
La particularité des données généralement disponibles en sécurité routière est qu'elles ne concernent que des données d'accidents, voire que des données d'accidents corporels ou mortels. Ainsi, l'effet causal de facteurs tels que l'alcool sur le risque d'accident ne peut être estimé, en raison de l'absence de données correspondant aux ≪ contrôles ≫ (les conducteurs non-impliqués dans un accident). Ces données sont ainsi soumises à un biais de sélection extrême.

Pour contourner ce problème, de nombreux travaux se sont concentrés sur l'estimation de mesures d'association entre certains facteurs et le risque d'être responsable d'un accident, parmi les conducteurs impliqués dans un accident (voire un accident corporel ou mortel). Un biais de sélection est donc toujours présent, et la question de l'interprétation causale des mesures d'association estimées n'a jamais été abordée, à notre connaissance. Les résultats de [Bareinboim and Pearl, 2012] et [Bareinboim and Tian, 2015] traitent de l'inférence causale en présence de biais de sélection. Ils sont introduits et illustrés sur des exemples simples dans le paragraphe 6.4.1. Leur application dans le cas des analyses en responsabilité est détaillée dans le paragraphe 6.4.2.

6.4.1 Inférence causale et biais de sélection

Le phénomène de biais de sélection a été largement étudié dans la littérature biostatistique [Robins, 2001, Hernán et al., 2004, Lajous et al., 2014], avec des exemples célèbres comme le biais de Berkson [Berkson, 1946]. Nous adoptons ici la terminologie utilisée dans [Bareinboim and Pearl, 2012, Bareinboim and Tian, 2015] où le biais de sélection est défini comme l'inclusion préférentielle de certains individus de la population. Afin de décrire cette sélection, une première étape consiste à introduire la variable binaire S qui indique

FIGURE 6.2 – Exemples de DAGs en présence de biais de sélection



l'inclusion dans l'étude. L'ajout de cette variable dans le DAG G conduit à un nouveau DAG G_s. Dans celui-ci, S est classiquement représentée de manière spécifique : S n'agit pas dans le modèle causal décrit par G mais joue un rôle sur le processus de sélection et les seules données disponibles sont celles pour lesquelles $S = 1$. Si S ne dépend d'aucune variable du DAG G, il n'y a aucune arête pointant vers S dans G_s, et donc pas de biais de sélection. Si, par contre, la sélection dans l'étude dépend de certaines variables V du DAG G, alors des arêtes pointent de V vers S dans G_s et on est en présence de biais de sélection. En fonction de la structure de G_s, ce biais de sélection peut conduire à des biais dans l'estimation des effets causaux [Hernán et al., 2004]. Un traitement complet de l'inférence causale en présence de biais de sélection est fourni dans les travaux de [Bareinboim and Pearl, 2012, Bareinboim and Tian, 2015], dont les résultats principaux sont illustrés sur des exemples simples ci-dessous. Motivé par le type de biais de sélection présent dans les analyses en responsabilité, nous nous concentrerons principalement sur les situations où la sélection dépend directement de la variable réponse Y, comme dans les DAGs 4 et 5 de la figure 6.2. A noter que les études cas-contrôles, classiques en recherche biomédicale, peuvent être vues comme des cas particuliers du DAG 4, voire du DAG 5 si l'inclusion dans l'étude dépend non seulement du statut par rapport à Y mais aussi de l'exposition X. Nous considérerons également le cas où S dépend d'un médiateur M entre X et Y, comme dans le DAG 6 de la figure 6.2.

En présence de biais de sélection, une première question naturelle est de savoir si l'effet causal de X sur Y est identifiable. Suivant la définition 2 de [Bareinboim and Tian, 2015], la loi $\mathbb{P}(Y(x) = y)$, pour $y \in \{0, 1\}$, est dite identifiable à partir de données souffrant de biais de sélection si les hypothèses induites par la structure du DAG G_s, composé des variables observées $\boldsymbol{\mathcal{V}}$ et S, la rendent exprimable à partir de la loi conditionnelle de $\boldsymbol{\mathcal{V}}|S = 1$. L'identifiabilité de $\mathbb{P}(Y(x) = y)$, pour tout x, est suffisante pour l'identifiabilité de l'excès de risque causal, mais aussi du risque relatif causal et de l'odds-ratio causal. Cependant, et comme nous le verrons ci-dessous, l'odds-ratio causal est identifiable dans certains cas où la loi $\mathbb{P}(Y(x) = y)$ ne l'est pas [Bareinboim and Pearl, 2012]. Lorsque ni $\mathbb{P}(Y(x) = y)$ ni l'odds-ratio causal n'est identifiable, des effets causaux alternatifs, tenant compte du conditionnement sur S d'une certaine manière, peuvent être estimés. Comme nous le verrons, la question de leur interprétation est liée à celle de la validité interne et externe.

Identifiabilité de $\mathbb{P}(Y(x) = y)$ en présence de biais de sélection

Notons comme précédemment G le DAG d'intérêt, composé des variables observées $\boldsymbol{\mathcal{V}}$, et par G_s le DAG obtenu après l'ajout de la variable S. Pour tout sous-ensemble de variables $\mathbf{C} \subseteq \boldsymbol{\mathcal{V}}$, on note $G_{\mathbf{C}}$ le sous-graphe de G restreint aux variables de $\mathbf{C}$. Pour tout $V_i \in \boldsymbol{\mathcal{V}}$, on note par ailleurs $An(V_i)_G$ l'union de V_i et de ses ancêtres dans le DAG G. Le théorème 2 de [Bareinboim and Tian, 2015] établit alors que $\mathbb{P}(Y(x) = y)$ est identifiable si et seulement si

$$An(Y)_{G_{\boldsymbol{\mathcal{V}}\setminus X}} \cap An(S)_{G_s} = \emptyset. \tag{R.1}$$

La condition (R.1) n'est clairement pas vérifiée dans les DAGs 4 et 5 où Y est un ancêtre de S. Plus précisément, on a par exemple $Y \in An(Y)_{G_{\boldsymbol{\mathcal{V}}\setminus X}} \cap An(S)_{G_s}$. Elle n'est pas non plus vérifiée dans le DAG 6 où $M \in An(Y)_{G_{\boldsymbol{\mathcal{V}}\setminus X}} \cap An(S)_{G_s}$.

Identifiabilité de l'odds-ratio causal en présence de biais de sélection

[Bareinboim and Pearl, 2012] introduisent la notion d'identifiabilité des odds-ratios conditionnels en présence de biais de sélection. Pour simplifier, nous considérerons ici l'identifiabilité des seuls odds-ratios conditionnels de la forme

$$OR(X, Y|\mathbf{W}) = \frac{\mathbb{P}(Y = 1|X = 1, \mathbf{W})/\mathbb{P}(Y = 0|X = 1, \mathbf{W})}{\mathbb{P}(Y = 1|X = 0, \mathbf{W})/\mathbb{P}(Y = 0|X = 0, \mathbf{W})},$$

où $\mathbf{W}$ est un vecteur de facteurs de confusion entre X et Y, comme dans les DAGs 4, 5 et 6. Par hypothèse, $\mathbf{W}$ contient l'ensemble des facteurs de confusion, et $OR(X, Y|\mathbf{W} = \mathbf{w})$ correspond donc à l'odds-ratio causal $\mathbf{W}$-spécifique

$$COR(X, Y|\mathbf{W}) = \frac{\mathbb{P}(Y(1) = 1|\mathbf{W})/\mathbb{P}(Y(1) = 0|\mathbf{W})}{\mathbb{P}(Y(0) = 1|\mathbf{W})/\mathbb{P}(Y(0) = 0|\mathbf{W})}.$$

D'après la définition 2 de [Bareinboim and Pearl, 2012], $OR(X, Y|\mathbf{W})$ est identifiable en présence de biais de sélection si les hypothèses induites par la structure du DAG le rendent exprimables en fonction de la distribution de $\boldsymbol{\mathcal{V}}|S = 1$. La symétrie de l'odds-ratio,

$$OR(X, Y|\mathbf{W}) = OR(Y, X|\mathbf{W}),$$

le rend identifiable dans certaines situations où $\mathbb{P}(Y(x) = y)$ ne l'est pas. Plus précisément, le théorème 1 de [Bareinboim and Pearl, 2012] établit que $OR(X, Y|\mathbf{W})$ est identifiable en présence de biais de sélection si et seulement si

$$X \perp\!\!\!\perp S|(Y, \mathbf{W}) \quad \text{ou} \quad Y \perp\!\!\!\perp S|(X, \mathbf{W}). \tag{R.2}$$

Cette condition est vérifiée sous le DAG 4, mais n'est pas garantie sous les DAGs 5 et 6. Ainsi, $OR(X, Y|\mathbf{W})$ est identifiable sous le DAG 4, mais ne l'est pas sous les DAGs 5 et 6.

Identifiabilité d'autres effets causaux en présence de biais de sélection

En résumé, les résultats de [Bareinboim and Tian, 2015] établissent que $\mathbb{P}(Y(x) = y)$ n'est pas identifiable sous les DAGs 4, 5 et 6 et donc que ni le risque relatif causal ni l'excès de risque causal ne peut généralement être estimé sous les DAGs de la figure 6.2. Dans le cas du DAG 4, les résultats de [Bareinboim and Pearl, 2012] établissent cependant que l'odds-ratio causal **w**-spécifique peut être estimé sans biais. Sous le DAG 4, la quantité $OR(X, Y|\mathbf{W} = \mathbf{w}, S = 1)$ est donc valide, de manière interne et externe [Kukull and Ganguli, 2012].

Sous les DAGs 5 et 6, l'odds-ratio causal **w**-spécifique ne peut pas être estimé sans hypothèses supplémentaires et la quantité $OR(X, Y|\mathbf{W} = \mathbf{w}, S = 1)$ n'est donc généralement pas valide de manière externe. La question de sa validité interne se pose alors naturellement. En fait, dans des situations telles que celles décrites par les DAGs 5 et 6, beaucoup d'épidemiologistes évoqueraient la présence d'un biais de sélection dans leur discussion, soulignant que leurs résultats ne décrivent que la population sélectionnée et ne sont peut-être pas généralisables à la population entière. Implicitement, ils suggéreraient ainsi la validité interne de leurs résultats (les odds-ratio conditionnels estimés sont des estimations correctes des effets causaux dans la population sélectionnée), et douteraient de leur validité externe (ces odds-ratio conditionnels estimés ne sont sans doute pas de bonnes estimations des effets causaux dans la population entière) [Kukull and Ganguli, 2012]. Ce type de raisonnement a conduit à des paradoxes célèbres dans la littérature, comme celui de l'obésité [2], qui peut être expliqué simplement par un phénomène de biais de sélection [Lajous et al., 2014]. Le cadre des SCMs introduit dans les paragraphes précédents est utile pour illustrer pourquoi des quantités telles que $OR(X, Y|\mathbf{W} = \mathbf{w}, S = 1)$ ne sont généralement pas valides, même de manière interne, sous les situations décrites par les DAGs 5 et 6.

Les DAGs 5 et 6 illustrent tous deux la situation où l'inclusion dans l'étude dépend d'un descendant de X, et est également liée à Y (S est soit un descendant de Y, soit un descendant d'un ancêtre de Y). En l'absence de biais de sélection, l'hypothèse d'ignorabilité conditionelle est vérifiée dans ces deux scénarios : $Y_x \perp\!\!\!\perp X|\mathbf{W}$. Mais sa version conditionnelle, sachant $S = 1$, n'est pas vérifiée en présence de biais de sélection. En fait, la variable S est ce que l'on appelle classiquement un « collider » : dans le DAG 5, elle dépend en particulier de X et Y, et dans le DAG 6, de M et U_S, même si la perturbation U_S n'est pas représentée sur la figure 6.2. Le conditionnement sur S peut alors induire des corrélations « artéfactuelles ». En particulier, on n'a généralement pas $U_X \perp\!\!\!\perp U_Y|S$, ce qui implique que l'on n'a généralement pas non plus $Y_x \perp\!\!\!\perp X|(\mathbf{W}, S)$. Ainsi, il n'est pas garanti que $\mathbb{P}(Y_x = 1|S = 1, \mathbf{W} = \mathbf{w}) = \mathbb{P}(Y = 1|X = x, S = 1, \mathbf{W} = \mathbf{w})$ sous les DAGs 5 et 6 (ni même sous le DAG 4). Par contre, sous ces DAGs on a toujours $Y_x \perp\!\!\!\perp X|(\mathbf{W}, S_x)$ et donc

$$\begin{aligned}\mathbb{P}(Y = 1|X = x, \mathbf{W} = \mathbf{w}, S = 1) &= \mathbb{P}(Y_x = 1|X = x, \mathbf{W} = \mathbf{w}, S_x = 1) \\ &= \mathbb{P}(Y_x = 1|\mathbf{W} = \mathbf{w}, S_x = 1).\end{aligned}$$

2. De nombreux résultats de la littérature établissent que l'obésité est un facteur protecteur du décès précoce chez les patients souffrant de maladies chroniques (telles que le diabète ou les maladies cardiovasculaires). Ceci a même conduit à ne pas recommander de perdre du poids aux patients obèses souffrant de ces maladies chroniques. Or ce paradoxe est sans doute faux ; il peut en tout cas s'expliquer simplement par le phénomène de biais de sélection [Lajous et al., 2014].

Considérons les deux groupes d'individus $\{S_x = 1, \mathbf{W} = \mathbf{w}\}$ pour $x \in \{0, 1\}$. Ils correspondent aux individus de la strate définie par $\mathbf{W} = \mathbf{w}$ qui auraient été sélectionnés dans le monde contrefactuel qui aurait suivi l'intervention $X = x$. Il est clair que les deux groupes $\{S_0 = 1, \mathbf{W} = \mathbf{w}\}$ et $\{S_1 = 1, \mathbf{W} = \mathbf{w}\}$ peuvent être relativement différents. Prenons l'exemple du risque relatif $\mathbb{P}(Y = 1|X = 1, \mathbf{W} = \mathbf{w}, S = 1)/\mathbb{P}(Y = 1|X = 0, \mathbf{W} = \mathbf{w}, S = 1)$, qui est égal à $\mathbb{P}(Y_1 = 1|\mathbf{W} = \mathbf{w}, S_1 = 1)/\mathbb{P}(Y_0 = 1|\mathbf{W} = \mathbf{w}, S_0 = 1)$. Parce que les groupes $\{S_0 = 1, \mathbf{W} = \mathbf{w}\}$ et $\{S_1 = 1, \mathbf{W} = \mathbf{w}\}$ sont composés d'individus différents, l'interprétation de cette quantité est délicate. Dans les situations décrites par les DAGs 4, 5 et 6, ce risque relatif n'est généralement valide ni de manière externe (il n'est pas égal à $\mathbb{P}(Y_1 = 1|\mathbf{W} = \mathbf{w})/\mathbb{P}(Y_0 = 1|\mathbf{W} = \mathbf{w})$), ni même de manière interne (il n'est pas non plus égal à $\mathbb{P}(Y_1 = 1|\mathbf{W} = \mathbf{w}, S = 1)/\mathbb{P}(Y_0 = 1|\mathbf{W} = \mathbf{w}, S = 1)$). Il en est de même pour l'excès de risque. Rappelons que sous le DAG 4, l'odds-ratio est valide de manière externe (et donc interne) puisque $OR(X, Y|\mathbf{W} = \mathbf{w}, S) = OR(X, Y|\mathbf{W} = \mathbf{w})$. Il n'est par contre généralement valide ni de manière externe ni de manière interne sous les DAGs 5 et 6.

6.4.2 Application aux analyses en responsabilité

Nous avons appliqué les principes exposés ci-dessus pour déterminer si l'effet causal de l'alcool sur le risque d'être responsable d'un accident est identifiable, à partir des données du projet ANR VOIESUR [3]. Les données de ce projet décrivent l'ensemble des accidents mortels et 5% des accidents corporels survenus en France en 2011. Dans ce projet, des experts ont analysé les rapports remplis par les forces de l'ordre suite à chacun des accidents pour évaluer la responsabilité des usagers impliqués dans ces accidents. Pour schématiser, un conducteur est jugé responsable par les experts s'il a, selon eux, déclenché l'accident, typiquement par une erreur ou une défaillance « coupable » (circulation en sens interdit, non respect d'un feu tricolore, absence de freinage, etc.). En d'autres termes, la responsabilité est une mesure (potentiellement entachée d'erreur) de la variable Faute, que l'on notera F. Cette variable est binaire et indique si le conducteur a commis une erreur, qui n'est en elle-même pas suffisante pour mener à un accident, mais serait considérée comme nécessaire dans la survenue de l'accident, compte tenu du contexte de l'accident, si ce dernier avait lieu. Dans le lexique causal, la variable Faute représente la présence d'une condition INUS, en lien avec une action, ou inaction, du conducteur. Cette variable est définie pour tous les conducteurs, pas seulement ceux impliqués dans un accident.

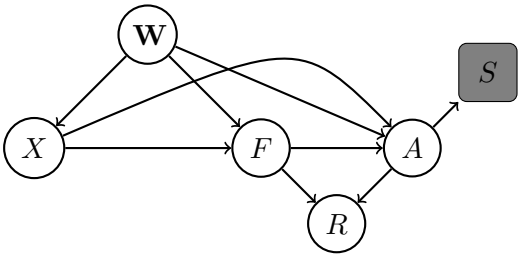
Pour simplifier l'exposé, nous nous concentrerons sur l'étude de l'effet causal de l'alcool sur le risque d'être responsable d'un accident mortel. Nous noterons alors A la variable binaire indiquant la survenue d'un accident mortel, et par X la variable binaire indiquant une alcoolémie supérieure au seuil légal. Nous considérerons également que seules les données relatives aux accidents mortels sont disponibles : un conducteur est donc inclus dans l'analyse seulement si $A = 1$ et S dépend donc de la variable A (on pourrait même considérer que $S = A$ puisque les données de VOIESUR sont censées renfermer l'ensemble des accidents mortels).

On peut définir une autre variable d'intérêt, que l'on notera R, et qui indique si le

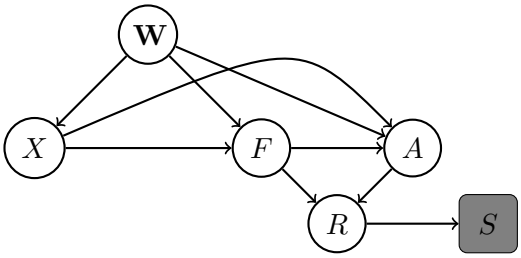
3. www.agence-nationale-recherche.fr/projet-anr/ ?tx_lwmsuivibilan_pi2[CODE]=ANR-11-VPTT-0007

FIGURE 6.3 – DAGs correspondant aux données disponibles et aux données « idéales » dans les analyses en responsabilité

Données disponibles



Données « idéales »



conducteur est responsable d'un accident mortel. Elle se définit donc par la relation $R = F \times A$. On a bien sûr, $R = 1$ si et seulement si $A = 1$ et $F = 1$. D'autre part, on a $R = 0$ si $A = 0$, même si $F = 1$, *i.e.*, même si le conducteur a commis une faute qui aurait amené les experts à le juger responsable de l'accident mortel si ce dernier était survenu.

On peut représenter l'ensemble de ces variables par le DAG de gauche de la figure 6.3. Le vecteur $\mathbf{W}$ représente l'ensemble des facteurs de confusion, que l'on supposera tous observés pour simplifier. Une remarque préalable est que sous le DAG de gauche de la figure 6.3, les résultats précédents établissent que $OR(X, A|\mathbf{W})$ est identifiable à partir de $OR(X, A|\mathbf{W}, S = 1)$. Or comme $S = 1 \Rightarrow A = 1$, on ne peut pas calculer $OR(X, A|\mathbf{W}, S = 1)$: l'effet causal de X sur A ne peut pas être estimé à partir de ces données, et on doit donc se restreindre aux effets de X sur F ou R. Deuxièmement, une arête particulièrement importante dans ce DAG est celle reliant X à A : celle-ci indique que $A \not\perp\!\!\!\perp X|F$, *i.e.*, le sur-risque d'accident mortel en lien avec l'alcool n'est pas entièrement « médié » par la variable F. Cette hypothèse semble naturelle puisque l'alcoolémie du conducteur est liée, sans doute de manière causale, à sa vitesse, qui est elle-même une cause de la gravité de l'accident.

Une application simple des principes présentés au paragraphe 6.4.1 permet d'établir que ni les lois $\mathbb{P}(R(x) = y)$ ou $\mathbb{P}(F(x) = y)$ pour $y \in \{0, 1\}$, ni les odds-ratios $OR(X, F|\mathbf{W})$ ou $OR(X, R|\mathbf{W})$ ne sont identifiables avec les données disponibles. On peut par contre montrer que si $A \perp\!\!\!\perp X|F$ dans le DAG précédent, alors on peut identifier $OR(X, F|\mathbf{W})$, et par suite approcher $OR(X, R|\mathbf{W})$. Par exemple, si des données étaient disponibles pour tous les types d'accidents (indépendamment de leur gravité), l'odds-ratio causal de l'alcool sur le risque d'être responsable d'un accident (quelqu'en soit sa gravité) pourrait être approché. Lorsque seules des données relatives aux accidents les plus graves sont disponibles, comme c'est le plus souvent le cas, deux stratégies sont envisageables. Reprenant l'exemple des accidents mortels, on peut premièrement chercher à modifier l'échantillon de départ. Le but est d'obtenir un échantillon proche d'une véritable étude cas-témoin, comparant des responsables d'accident mortel ($R = 1$), et des conducteurs non-responsables d'accident mortel ($R = 0$), comme dans le DAG de droite de la figure 6.3. En particulier, la population des contrôles est principalement composée de conducteurs tels que $\{F = 0, A = 0\}$, alors que

l'échantillon initial est composé uniquement de conducteurs pour lesquels $\{F = 0, A = 1\}$. Comme $A \not\perp X|F$, le groupe contrôle initial est typiquement différent du groupe contrôle théorique en terme d'exposition au facteur X. Certaines modifications ont été proposées dans la littérature [Laumon et al., 2005]. Cependant, la limite principale de cette approche réside dans l'impossibilité de tester si l'échantillon finalement obtenu peut être décrit par le DAG de droite de la figure 6.3. Une recommandation pourrait être d'envisager plusieurs transformations de l'échantillon initial, et vérifier la robustesse des résultats par des analyses de sensibilité. Une autre solution consiste à s'abstenir d'estimer les odds-ratios causaux du type $COR(X, R|\mathbf{W})$ et de se limiter aux effets causaux tels que $\mathbb{P}(R_1 = 1|\mathbf{W} = \mathbf{w}, S_1 = 1)/\mathbb{P}(R_0 = 1|\mathbf{W} = \mathbf{w}, S_0 = 1)$. La principale limite de cette stratégie réside dans la difficulté d'interprétation de ces quantités.

6.4.3 Discussion

Quelque soit l'approche retenue, d'autres biais et difficultés sont à considérer en vue de l'estimation des effets causaux dans les analyses en responsabilité. En effet, comme évoqué plus haut, l'effet causal n'est en général pas identifiable si des facteurs de confusion ne sont pas observés. Or, l'ensemble des facteurs de confusion n'est jamais observé dans les études épidémiologiques. En sécurité routière par exemple, des variables telles que l'usage du téléphone au volant, la prise de médicament ou encore le goût du risque ne sont généralement pas mesurées. Une autre difficulté vient du fait que certains de nos témoins sont appariés sur les cas (les non-responsables d'un accident impliquant deux véhicules typiquement), alors que d'autres témoins ne le sont pas (les non-responsables d'accident à un véhicule). Une autre difficulté est la présence de données manquantes (notamment pour la variable vitesse, avec un mécanisme de données manquantes qui n'est pas nécessairement aléatoire). Enfin, une dernière source de biais notable concerne la qualité de la mesure de la variable Faute, faite par les experts. En particulier, si les erreurs de mesure dépendent de certaines variables (par exemple, si les experts jugent la responsabilité des conducteurs de manière différente en fonction de l'alcoolémie du conducteur), alors de nouveaux biais surviennent. La question de la validité de la détermination de la responsabilité par les experts fait l'objet d'un stage de M2 qui débute ce printemps.

En résumé, même si la nature causale de l'effet de certains facteurs (alcool, vitesse, etc.) sur le risque d'accident et le risque d'être responsable d'un accident est communément admise, la quantification de ces effets reste une question délicate à partir des données disponibles.

6.5 Autres perspectives : causalité et grande dimension

La connaissance de la structure du DAG, qui décrit le modèle causal d'intérêt, est primordiale pour l'inférence causale. Dans un contexte où l'on considère un nombre restreint de variables dans ce modèle (au risque d'omettre certaines variables importantes, et d'invalider l'hypothèse d'ignorabilité conditionnelle), des connaissances « expertes » peuvent permettre la construction du DAG à la main. Par contre, dans un contexte de grande

dimension où le nombre de variables est élevé, cette construction à la main n'est plus envisageable. En particulier, les données décrivant les prescriptions médicamenteuses ont été récoltées par une équipe d'épidémiologistes de Bordeaux (sous la direction d'Emmanuel Lagarde, INSERM). Dans l'optique d'estimer les effets causaux de différentes classes de médicaments sur le risque d'être responsable d'un accident de la route, on pourra avoir recours aux approches présentées par exemple dans [Kalisch et al., 2012] pour inférer la structure du DAG à partir de données observationnelles.

D'autre part, une prépublication récente [Bloniarz et al., 2015] décrit l'intérêt du lasso pour estimer l'effet causal d'un traitement dans le cadre de l'essai thérapeutique (et donc dans le cadre de données interventionnelles et non pas observationnelles). Dans l'essai thérapeutique, le traitement est randomisé, ce qui implique que l'hypothèse d'ignorabilité $(Y(0), Y(1)) \perp\!\!\!\perp X$ est en principe vérifiée et aucun facteur de confusion ne vient perturber l'estimation de l'effet causal de X sur Y. Dans [Bloniarz et al., 2015], les auteurs montrent cependant que l'ajustement sur des covariables disponibles permet d'améliorer la précision de l'estimation. Lorsque ces covariables sont nombreuses, l'utilisation du lasso se révèle être une stratégie adaptée, afin notamment d'identifier les interactions entre le type de traitement et les covariables. Dans [Bloniarz et al., 2015], les auteurs considèrent le cas classique d'un essai thérapeutique où le type de traitement est binaire : placebo ou nouveau traitement par exemple. Dans le cas d'essais thérapeutiques à plusieurs bras, AutoRefLasso ou CliqueFused, décrites au paragraphe 4.3, pourraient être envisagées afin d'identifier les interactions entre le type de traitement et les covariables.

Bibliographie Vivian Viallon (2009-2016)

Texte

[VV1] Nada Assi, Anne Fages, Paolo Vineis, Marc Chadeau-Hyam, Magdalena Stepien, Talita Duarte-Salles, Graham Byrnes, Houda Boumaza, Sven Knüppel, Tilman Kühn, Domenico Palli, Christina Bamia, Hendriek Boshuizen, Catalina Bonet, Kim Overvad, Mattias Johansson, Ruth Travis, Marc Gunter, Eiliv Lund, Laure Dossus, Bénédicte Elena-Herrmann, Elio Riboli, Mazda Jenab, Vivian Viallon, and Pietro Ferrari. A statistical framework to model the meeting-in-the-middle principle using metabolomic data : application to hepatocellular carcinoma in the EPIC study. *Mutagenesis*, 30(6) :743–753, 2015.

[VV2] Paul Blanche, Aurélien Latouche, and Vivian Viallon. Time-dependent AUC with right-censored data : A survey. In *Risk Assessment and Evaluation of Predictions*, pages 239–251. Springer, 2013.

[VV3] Joël Coste, Frédérique Tissier, Jacques Pouchot, Emmanuel Ecosse, Alexandra Rouquette, Xavier Bertagna, Rossella Libé, and Vivian Viallon. Rasch analysis for assessing unidimensionality and identifying measurement biases of malignancy scores in oncology. the example of the Weiss histopathological system for the diagnosis of adrenocortical cancer. *Cancer Epidemiology*, 38(2) :200–208, 2014.

[VV4] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8(4) :667–698, 2012.

[VV5] Charly Empereur-mot, Hélène Guillemain, Aurélien Latouche, Jean-François Zagury, Vivian Viallon, and Matthieu Montes. Predictiveness curves in virtual screening. *Journal of Cheminformatics*, 7(1) :1–17, 2015.

[VV6] Thomas Lieutaud, Amina Ndiaye, Mireille Chiron, Blandine Gadegbeku, and Vivian Viallon. The epidemiology of traumatic brain injury deriving from road traffic collision : trend changes following strengthened legislative measures in france. Soumis, 2015.

[VV7] Edouard Ollier, Adeline Samson, Xavier Delavenne, and Vivian Viallon. A SAEM algorithm for fused lasso penalized non linear mixed effect models : Application to group comparison in pharmacokinetic. *Computational Statistics and Data Analysis*, A paraître, 2015.

[VV8] Edouard Ollier and Vivian Viallon. Regression modeling on stratified data : automatic and covariate-specific selection of the reference stratum with simple l_1-norm penalties. *arXiv preprint arXiv :1508.05476*, 2015.

[VV9] Vivian Viallon, Onureena Banerjee, Eric Jougla, Grégoire Rey, and Joel Coste. Empirical comparison study of approximate methods for structure selection in binary graphical models. *Biometrical Journal*, 56(2) :307–331, 2014.

[VV10] Vivian Viallon, Emmanuel Ecosse, Mounir Mesbah, Jacques Pouchot, and Joël Coste. Using extended Rasch models to assess validity of diagnostic tests in the presence of a reference standard. *Journal of Applied Measurement*, 13(4) :376–393, 2011.

[VV11] Vivian Viallon, Sophie Lambert-Lacroix, Hölger Hoefling, and Franck Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26(1) :285–301, 2016.

[VV12] Vivian Viallon and Aurélien Latouche. Discrimination measures for survival outcomes : connection between the auc and the predictiveness curve. *Biometrical Journal*, 53(2) :217–236, 2011.

[VV13] Vivian Viallon and Bernard Laumon. Fractions of fatal crashes attributable to speeding : Evolution for the period 2001–2010 in France. *Accident Analysis & Prevention*, 52 :250–256, 2013.

[VV14] Vivian Viallon, Stéphane Ragusa, Françoise Clavel-Chapelon, and Jacques Bénichou. How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine*, 28(6) :901–916, March 2009.

Travaux antérieurs

Travaux en statistique mathématique

[7] B. Maillot et V. Viallon. *Uniform limit laws of the logarithm for nonparametric estimators of the regression function in presence of censored data.* Mathematical Methods of Statistics, 18(2) :159-184 (2009).

[6] M. Debbarh, V. Viallon. *Testing additivity in nonparametric regression under random censorship.* Stat. and Prob. Letters, 78(16) :2584-2591 (2008).

[5] M. Debbarh, V. Viallon. *Uniform limit laws of the logarithm for the additive regression function in presence of censored data.* Electronic Journal of Statistics, 2 : 516-541 (2008).

[4] V. Viallon. *Uniform law of the logarithm for a nonparametric estimator of the regression function in the presence of censored data.* C. R. Acad. Sci. Paris, Ser. I, 346(4) : 225-228 (2008).

[3] V. Viallon. *Functional limit laws for the increments of the quantile process ; with applications.* Electronic Journal of Statistics, 1 : 496-518 (2007).

[2] M. Debbarh, V. Viallon. *Uniform convergence for estimators of the additive regression function under random censorship.* C. R., Math., Acad. Sci. Paris, Ser. I, 345(2) : 97-100 (2007).

[1] M. Debbarh, V. Viallon. *Mean square convergence for estimators of additive regression under random censorship.* C. R. Acad. Sci. Paris, Ser. I, 344(3) : 205-210 (2007).

Travaux appliqués

[17] N. Assi, A. Moskal, N. Slimani, V. Viallon, V. Chajes, (...), I. Romieu, P. Ferrari. *A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition study.* Public Health Nutr., 2015.

[16] A. Grasset, V. Viallon, E. Amoros, M. Hours. *Typology of bicycle crashes based on a survey of a thousand of injured cyclists from a road trauma registry.* Advances in Transportation Studies, 2 (Special Issue), 17-28 (2014).

[15] F. Stenard, O. Morales, K. Ghazal, V. Viallon, (...), F. Conti. *CD49b, a major marker of regulatory T-cells type 1, predict the response to antiviral therapy of recurrent hepatitis C after liver transplantation.* Biomed Res. International ; 2014 :290878 (2014).

[14] A. Hüsing, F. Canzian, L. Beckmann, M. Garcia-Closas, W.R. Diver, M.J. Thun, C.D. Berg, R.N. Hoover, R.G. Ziegler, J.D. Figueroa, C. Isaacs, A. Olsen, V. Viallon, H. Boeing, (...), R. Kaaks ; on behalf of the BPC3. *Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status.* J. Med. Genet. ;49(9) :601-608 (2012).

[13] F. Tissier, S. Aubert, E. Leteurtre, A. Al Ghuzlan, M. Patey, M. Decaussin, L. Doucet, F. Gobet, C. Hoang, C. Mazerolles, G. Monges, K. Renaudin, N. Sturm, H. Trouette, M.C. Vacher-Lavenu, V. Viallon, E. Baudin, X. Bertagna, J. Coste, R. Libe. *Adrenocortical tumors : improving the practice of the Weiss system through virtual microscopy : a National Program of the French Network INCa-COMETE.* Am. J. Surg. Pathol. 36(8) :1194-201 (2012).

[12] F. Campeotto, A. Suau, N. Kapel, F. Magne, V. Viallon, L. Ferraris, A. J. Waligora-Dupriet, P. Soulaines, B. Leroux, N. Kalach, C. Dupont, M. J. Butel. *A fermented formula in pre-term infants : clinical tolerance, gut microbiota, down-regulation of faecal calprotectin and up-regulation of faecal secretory IgA.* British Journal of Nutrition. 22 :1-10 (2011).

[11] C. Espy, W. Morelle, N. Kavian, P. Grange, C. Goulvestre, V. Viallon, C. Chereau, C. Pagnoux, J.C. Michalski, L. Guillevin, B. Weill, F. Batteux, P. Guilpain. *Sialylation level of anti-proteinase 3 (PR3) antibodies determines the activity of Wegener's granulomatosis.* Arthritis Rheum. 63(7) :2105-2115 (2011).

[10] J. Toubiana, E. Courtine, F. Pène, V. Viallon, P. Asfar, C. Daubin, C. Rousseau, C. Chenot, F. Ouazz, D. Grimaldi, A. Cariou, J.D. Chiche, J.P. Mira. *IRAK1 variant and septic shock.* Crit. Care Med. 38(12) :2287-94 (2010).

[9] E. Frisan, P. Pawlikowska, C. Pierre-Eugène, V. Bardet, V. Viallon, L. Gibault, O. Kosmider, S. Park, F. Kuhnowsky, M. Guesnu, P. Mayeux, C. Lacombe, F. Dreyfus, F. Porteu, M. Fontenay. *p-ERK1/2 is a predictive factor of response to erythropoiesis-stimulating agents in low/int-1 myelodysplastic syndromes.* Haematologica. 95(11) :1964-1968 (2010).

[8] C. Cartry, V. Viallon, P. Hornoy, C. Adamsbaum. *Diffusion-weighted MR imaging of the normal fetal brain : marker of fetal brain maturation.* J. Radiol., 91 : 561-566 (2010).

[7] C. Chaussain-Miller, S. Opsahl-Vital, V. Viallon, L. Vermelin, M. Sixou, J.J. Lasfargues. *Predictive performance of a new caries test for patients undergoing orthodontic treatment.* Clinical Oral Investigations, 14(2) :177-185 (2009).

[6] P. Fauque, P. Jouannet, C. Davy, J. Guibert, V. Viallon, S. Epelboin, J.M. Kunstmann, C. Patrat. *Cumulative results including obstetrical and neonatal outcome of fresh and frozen-thawed cycles in elective single versus double fresh embryo transfers.* Fertility and Sterility, 94(3) : 927-35 (2009).

[5] P. Fauque, M. Albert, C. Serres, V. Viallon, C. Chalas, C. Davy, S. Epelboin, P. Jouannet, C. Patrat. *From ultrastructural flagellar sperm defects to health of babies*

conceived by ICSI Reproductive BioMedicine Online, 19(3) : 326-36 (2009).

[4] F. Campeotto, M. Baldassare, M.J. Butel, V. Viallon, F. Nganzali, P. Soulaines, N. Kalach, A. Lapillone, N. Laforgia, G. Moriette, C. Dupont, N. Kapel. *Fecal calprotectin as a noninvasive marker of digestive distress in preterm neonates : cut-off levels.* Journal of Pediatric Gastroenterology and Nutrition, 48(4) : 507-10 (2009).

[3] C. Patrat, I. Okamoto, P. Diabangouaya, V. Viallon, P. Le Baccon, E. Heard. *Dynamic changes in paternal X-chromosome activity during imprinted X inactivation in mice.* PNAS, 106(13) : 5198-203 (2009).

[2] C.B. d'Alva, G. Abiven-Leplace, V. Viallon, X. Bertagna, J. Bertherat. *Sex steroids in androgen-secreting adrenocortical tumors : clinical and hormonal features in comparison with non tumoral causes of androgen excess.* European Journal of Endocrinology, 159(5) :641-647 (2008).

[1] F. Pène, S. Percheron, V. Lemiale, V. Viallon, Y.E. Claessens, S. Marqué, J. Charpentier, D.C. Angus, A. Cariou, J.D. Chiche and J.P. Mira. *Temporal changes in management and outcome of septic shock in patients with malignancies in the intensive care unit.* Critical Care Medicine, 36(3) : 690-696 (2008).

Bibliographie

[Aalen et al., 2008] Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis : a process point of view*. Springer Science & Business Media.

[Ahmed and Xing, 2009] Ahmed, A. and Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29) :11878–11883.

[Anandkumar et al., 2012] Anandkumar, A., Tan, V. Y., Huang, F., Willsky, A. S., et al. (2012). High-dimensional structure estimation in Ising models : Local separation criterion. *The Annals of Statistics*, 40(3) :1346–1375.

[Andersen et al., 2012] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

[Argyriou et al., 2008] Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3) :243–272.

[Bach et al., 2012] Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012). Structured sparsity through convex optimization. *Statistical Science*, 27(4) :450–468.

[Bach, 2008] Bach, F. R. (2008). Bolasso : model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

[Banerjee et al., 2008] Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9 :485–516.

[Barber and Candès, 2015] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5) :2055–2085.

[Bareinboim and Pearl, 2012] Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012) ; JMLR*, 22 :100–108.

[Bareinboim and Tian, 2015] Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI*, pages 3475–3481.

[Becker et al., 2011] Becker, S. R., Candès, E. J., and Grant, M. C. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3) :165–218.

[Bell, 1934] Bell, E. T. (1934). Exponential numbers. *American Mathematical Monthly*, pages 411–419.

[Berkson, 1946] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2 :47–53.

[Besag, 1975] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, pages 179–195.

[Beyersmann et al., 2011] Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing risks and multistate models with R*. Springer Science & Business Media.

[Bickel et al., 2009] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

[Bloniarz et al., 2015] Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J., and Yu, B. (2015). Lasso adjustments of treatment effect estimates in randomized experiments. *arXiv preprint arXiv :1507.03652*.

[Bonnefoy et al., 2014] Bonnefoy, A., Emiya, V., Ralaivola, L., and Gribonval, R. (2014). A dynamic screening principle for the lasso. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 6–10. IEEE.

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1) :1–122.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

[Brubacher et al., 2014] Brubacher, J., Chan, H., and Asbridge, M. (2014). Culpability analysis is still a valuable technique. *International Journal of Epidemiology*, 43(1) :270–272.

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data : methods, theory and applications*. Springer Science & Business Media.

[Buyse et al., 2006] Buyse, M., Loi, S., Van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., et al. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*, 98(17) :1183–1192.

[Candes and Tao, 2007] Candes, E. and Tao, T. (2007). The dantzig selector : statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

[Candes et al., 2008] Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6) :877–905.

[Chadeau-Hyam et al., 2011] Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M., Jenab, M., Sacerdote, C., Bruce, S. J., Holmes, E., and Vineis, P. (2011). Meeting-in-the-middle using metabolic profiling–a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, 16(1) :83–88.

[Chambaz et al., 2014] Chambaz, A., Drouet, I., and Thalabard, J.-C. (2014). Causality, a trialogue. *Journal of Causal Inference*, 2(2) :201–241.

[Colditz et al., 2000] Colditz, G., Atwood, K., Emmons, K., Monson, R., Willett, W., Trichopoulos, D., and Hunter, D. (2000). Harvard report on cancer prevention volume 4 : Harvard cancer risk index. *Cancer causes & control*, 11(6) :477–488.

[Colditz et al., 2004] Colditz, G. A., Rosner, B. A., Chen, W. Y., Holmes, M. D., and Hankinson, S. E. (2004). Risk factors for breast cancer according to estrogen and progesterone receptor status. *Journal of the National Cancer Institute*, 96(3) :218–228.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

[Cox and Wermuth, 1994] Cox, D. R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika*, 81(2) :403–408.

[Dai and Pelckmans, 2012] Dai, L. and Pelckmans, K. (2012). An ellipsoid based, two-stage screening test for bpdn. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 654–658. IEEE.

[Dalalyan et al., 2014] Dalalyan, A. S., Hebiri, M., and Lederer, J. (2014). On the prediction performance of the lasso. *arXiv preprint arXiv :1402.1700*.

[Danaher et al., 2014] Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2) :373–397.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, pages 94–128.

[Dempster, 1972] Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

[Dezeure et al., 2014] Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2014). High-dimensional inference : Confidence intervals, p-values and r-software hdi. *arXiv preprint arXiv :1408.4026*.

[Donoho and Tsaig, 2008] Donoho, D. L. and Tsaig, Y. (2008). Fast solution of l_1-norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54(11) :4789–4812.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, 32 :407–499.

[Elvik et al., 2005] Elvik, R., Christensen, P., and Amundsen, A. H. (2005). Speed and road accidents : an evaluation of the power model. *Nordic Road and Transport Research*, 17(1).

[Evgeniou and Pontil, 2004] Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456) :1348–1360.

[Fan and Lv, 2008] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(5) :849–911.

[Fan and Lv, 2010] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica*, 20 :101–148.

[Fercoq et al., 2015] Fercoq, O., Gramfort, A., and Salmon, J. (2015). Mind the duality gap : safer rules for the lasso. *arXiv preprint arXiv :1505.03410*.

[Forman, 2003] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3 :1289–1305.

[Friedman et al., 2007] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, 1(2) :302–332.

[Friedman et al., 2008] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.*, 33(1) :1–22.

[Gail et al., 1989] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24) :1879–1886.

[Gail and Pfeiffer, 2005] Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2) :227–239.

[Gertheiss and Tutz, 2012] Gertheiss, J. and Tutz, G. (2012). Regularization and model selection with categorial effect modifiers. *Statistica Sinica*, 22 :957–982.

[Giraud, 2014] Giraud, C. (2014). *Introduction to high-dimensional statistics*. CRC Press.

[Goeman et al., 2012] Goeman, J., Meijer, R., and Chaturvedi, N. (2012). penalized : l_1 (lasso and fused lasso) and l_2 (ridge) penalized estimation in glms and in the cox model. *URL http ://cran. r-project. org/web/packages/penalized/index. html*.

[Greenland et al., 1999] Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48.

[Guo et al., 2015] Guo, J., Cheng, J., Levina, E., Michailidis, G., and Ji, Z. (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics*, 9(2) :821–848.

[Hamburg and Collins, 2010] Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4) :301–304.

[Hernán et al., 2004] Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5) :615–625.

[Hill, 1965] Hill, A. B. (1965). The environment and disease : association or causation ? *Proceedings of the Royal Society of Medicine*, 58(5) :295–300.

[Hocking, 1976] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, pages 1–49.

[Höfling et al., 2010] Höfling, H., Binder, H., and Schumacher, M. (2010). A coordinate-wise optimization algorithm for the Fused Lasso. *Arxiv preprint arXiv :1011.6409*.

[Höfling and Tibshirani, 2009] Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10 :883–906.

[Hume, 1739] Hume, D. (1739). Traité de la nature humaine.

[IARC, 2001] IARC (2001). *IARC monographs on the evaluation of carcinogenic risks to humans*, volume 78. International Agency for Research on Cancer.

[Imai et al., 2010] Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.

[Jojic et al., 2011] Jojic, V., Saria, S., and Koller, D. (2011). Convex envelopes of complexity controlling penalties : the case against premature envelopment. In *International Conference on Artificial Intelligence and Statistics*, pages 399–406.

[Kalbfleisch and Prentice, 2011] Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

[Kalisch et al., 2012] Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11) :1–26.

[Kannel et al., 1976] Kannel, W. B., McGee, D., and Gordon, T. (1976). A general cardiovascular risk profile : the framingham study. *The American journal of cardiology*, 38(1) :46–51.

[Kim et al., 2007] Kim, S.-J., Koh, K., Lustig, M., Boyd, S., and Gorinevsky, D. (2007). An interior-point method for large-scale l_1-regularized least squares. *IEEE J. Select. Top. Sign. Process.*, 1(4) :606–617.

[Kim et al., 2012] Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, 13(1) :1037–1057.

[Kouno et al., 2013] Kouno, T., de Hoon, M., Mar, J. C., Tomaru, Y., Kawano, M., Carninci, P., Suzuki, H., Hayashizaki, Y., and Shin, J. W. (2013). Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome Biology*, 14 :R118.

[Kukull and Ganguli, 2012] Kukull, W. A. and Ganguli, M. (2012). Generalizability. the trees, the forest, and the low-hanging fruit. *Neurology*, 78(23) :1886–1891.

[Lajous et al., 2014] Lajous, M., Bijon, A., Fagherazzi, G., Boutron-Ruault, M.-C., Balkau, B., Clavel-Chapelon, F., and Hernán, M. A. (2014). Body mass index, diabetes, and mortality in french women : explaining away a "paradox". *Epidemiology (Cambridge, Mass.)*, 25(1) :10.

[Laumon et al., 2005] Laumon, B., Gadegbeku, B., Martin, J.-L., and Biecheler, M.-B. (2005). Cannabis intoxication and fatal road crashes in france : population based case-control study. *Bmj*, 331(7529) :1371.

[Lauritzen, 1996] Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

[Lee et al., 2013] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013). Exact post-selection inference with the lasso. *arXiv preprint arXiv :1311.6238*.

[Lounici et al., 2011] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, pages 2164–2204.

[Lozano and Swirszcz, 2012] Lozano, A. C. and Swirszcz, G. (2012). Multi-level lasso for sparse multi-task regression. In *ICML*.

[Lunn and McNeil, 1995] Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, pages 524–532.

[Mackie, 1974] Mackie, J. L. (1974). *The cement of the universe : a study of causation*. Oxford, oxford university press edition.

[Maurer and Pontil, 2013] Maurer, A. and Pontil, M. (2013). Excess risk bounds for multitask learning with trace norm regularization. *JMLR, W& CP*, 30 :55–76.

[McCarthy et al., 2015] McCarthy, A., Keller, B., Kontos, D., Boghossian, L., McGuire, E., Bristol, M., Chen, J., Domchek, S., and Armstrong, K. (2015). The use of the gail model, body mass index and snps to predict breast cancer among women with abnormal (bi-rads 4) mammograms. *Breast Cancer Res*, 17(1) :1.

[Meinshausen, 2007] Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1) :374–393.

[Meinshausen and Bühlmann, 2006] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462.

[Meinshausen and Bühlmann, 2010] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(4) :417–473.

[Meinshausen et al., 2009] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104 :1671–1681.

[Mill, 1856] Mill, J. S. (1856). *A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles, and the Methods of Scientific Investigation*, volume 2. JW Parker.

[Munsell et al., 2014] Munsell, M. F., Sprague, B. L., Berry, D. A., Chisholm, G., and Trentham-Dietz, A. (2014). Body mass index and breast cancer risk according to post-menopausal estrogen-progestin use and hormone receptor status. *Epidemiologic reviews*, 36(1) :114–136.

[Ndiaye et al., 2015] Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. (2015). Gap safe screening rules for sparse multi-task and multi-class models. *arXiv preprint arXiv :1506.03736*.

[Negahban and Wainwright, 2011] Negahban, S. N. and Wainwright, M. J. (2011). Simultaneous support recovery in high dimensions : Benefits and perils of block-regularization. *Information Theory, IEEE Transactions on*, 57(6) :3841–3863.

[Nilsson, 2004] Nilsson, G. (2004). *Traffic safety dimensions and the power model to describe the effect of speed on safety*. PhD thesis, Lund University.

[Oelker et al., 2014] Oelker, M.-R., Gertheiss, J., and Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, 14(2) :157–177.

[Park and Hastie, 2007] Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 69(4) :659–677.

[Pearl, 1995] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4) :669–688.

[Pearl, 2000] Pearl, J. (2000). *Causality : models, reasoning, and inference*. Cambridge Univ Press.

[Pearl, 2009] Pearl, J. (2009). Causal inference in statistics : An overview. *Statistics Surveys*, 3 :96–146.

[Qian and Jia, 2016] Qian, J. and Jia, J. (2016). On pattern recovery of the fused lasso. *Computational Statistics & Data Analysis*, 94 :221–237.

[Ravikumar et al., 2010] Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional Ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38(3) :1287–1319.

[Reulen and Kneib, 2015] Reulen, H. and Kneib, T. (2015). Structured fusion lasso penalised multi-state models. Technical report, University of Goettingen.

[Robins, 1986] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9) :1393–1512.

[Robins, 2001] Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12(3) :313–320.

[Rosner et al., 2013] Rosner, B., Glynn, R. J., Tamimi, R. M., Chen, W. Y., Colditz, G. A., Willett, W. C., and Hankinson, S. E. (2013). Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *American Journal of Epidemiology*, 178(2) :296–308.

[Rothman et al., 2008] Rothman, K. J., Greenland, S., and Lash, T. L. (2008). Modern epidemiology. 3rd edition. *Philadephia : Lippincott Williams & Wilkins*.

[Rubin, 1974] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5) :688.

[Salmi et al., 2014] Salmi, L. R., Orriols, L., and Lagarde, E. (2014). Comparing responsible and non-responsible drivers to assess determinants of road traffic collisions : time to standardise and revisit. *Injury prevention*, 20 :380–386.

[Santhanam and Wainwright, 2012] Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *Information Theory, IEEE Transactions on*, 58(7) :4117–4134.

[Schwaller et al., 2015] Schwaller, L., Robin, S., and Stumpf, M. (2015). Bayesian inference of graphical model structures using trees.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

[Sharpnack et al., 2012] Sharpnack, J., Rinaldo, A., and Singh, A. (2012). Sparsistency of the edge lasso over graphs. *AISTAT*.

[She, 2010] She, Y. (2010). Sparse regression with exact clustering. *Electronic Jounal of Statistics*, 4 :1055–1096.

[Suzuki et al., 2009] Suzuki, R., Orsini, N., Saji, S., Key, T. J., and Wolk, A. (2009). Body weight and incidence of breast cancer defined by estrogen and progesterone receptor status—a meta-analysis. *International journal of cancer*, 124(3) :698–712.

[Tamimi et al., 2012] Tamimi, R. M., Colditz, G. A., Hazra, A., Baer, H. J., Hankinson, S. E., Rosner, B., Marotti, J., Connolly, J. L., Schnitt, S. J., and Collins, L. C. (2012). Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. *Breast cancer research and treatment*, 131(1) :159–167.

[Therneau and Grambsch, 2000] Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data : extending the Cox model*. Springer Science & Business Media.

[Tian and Pearl, 2002] Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference of Artificial Intelligence*, pages 567–573. AAAI Press/ The MIT Press, Menlo Park, CA.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 58 :267–288.

[Tibshirani et al., 2012] Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(2) :245–266.

[Tibshirani et al., 2005] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108.

[Tibshirani and Wang, 2008] Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1) :18–29.

[Tibshirani and Taylor, 2011] Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3) :1335–1371.

[Van de Geer et al., 2014] Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3) :1166–1202.

[van Houwelingen and Le Cessie, 1990] van Houwelingen, J. and Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in medicine*, 9(11) :1303–1325.

[Varoquaux et al., 2012] Varoquaux, G., Gramfort, A., and Thirion, B. (2012). Small-sample brain mapping : sparse recovery on spatially correlated designs with randomization and clustering. *arXiv preprint arXiv :1206.6447*.

[Wainwright, 2009] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5) :2183–2202.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2) :1–305.

[Wang et al., 2007] Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3) :553–568.

[Wang et al., 2013] Wang, J., Zhou, J., Wonka, P., and Ye, J. (2013). Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, pages 1070–1078.

[Wang et al., 2009] Wang, P., Chao, D. L., and Hsu, L. (2009). Learning networks from high dimensional binary data : An application to genomic instability data. *arXiv preprint arXiv :0908.3882*.

[Xiang and Ramadge, 2012] Xiang, Z. J. and Ramadge, P. J. (2012). Fast lasso screening tests based on correlations. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2137–2140. IEEE.

[Xiang et al., 2014] Xiang, Z. J., Wang, Y., and Ramadge, P. J. (2014). Screening tests for lasso problems. *arXiv preprint arXiv :1405.4897*.

[Xiang et al., 2011] Xiang, Z. J., Xu, H., and Ramadge, P. J. (2011). Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems*, pages 900–908.

[Yang and Ravikumar, 2011] Yang, E. and Ravikumar, P. K. (2011). On the use of variational inference for learning discrete graphical model. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1009–1016.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67.

[Zhang and Zhang, 2014] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(1) :217–242.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7 :2541–2563.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320.

Chapitre A

Principes généraux des approches pénalisées

Cette annexe a pour vocation d'introduire les principes généraux des approches d'estimation par pénalisation d'un risque empirique, et notamment du lasso [Tibshirani, 1996]. Par souci de lisibilité, elle reprend en grande partie les idées présentées dans la section 1.1.2 du chapitre introductif, en les complétant. On y présente également brièvement diverses extensions du lasso ainsi que des stratégies pour sélectionner les paramètres de régularisation en pratique. Pour simplifier l'exposé, nous nous plaçons dans le cas du modèle linéaire homoscédastique sur design déterministe, mais les principes s'étendent naturellement à une large variété de modèles paramétriques (modèles linéaires généralisés, etc.).

A.1 Le modèle de régression linéaire

On se place dans le cadre de l'étude de l'association entre une variable d'intérêt réelle et un vecteur de covariables. En notant $n \geq 1$ le nombre d'observations, nous supposons disposer d'une matrice de design déterministe $\mathbf{X} \in \mathbb{R}^{n \times p}$, avec $p \geq 1$. On notera $\mathbf{x}_i \in \mathbb{R}^p$ sa i-ème ligne et $X_j \in \mathbb{R}^n$ sa j-ème colonne. On suppose disposer par ailleurs d'un échantillon $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ de n observations d'une variable aléatoire d'intérêt, sous le modèle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \quad \forall i \in [n], \tag{A.1}$$

où, pour tout entier $m \geq 1$, $[m]$ désigne l'ensemble $\{1, \ldots, m\}$. Le vecteur $\boldsymbol{\beta}^* \in \mathbb{R}^p$ renferme les paramètres du modèle à estimer, et on supposera ici que $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ où les $(\varepsilon_i)_{i \in [n]}$ sont indépendants et identiquement distribués (*i.i.d.*), de loi normale $\mathcal{N}(0, \sigma^2)$ avec $\sigma > 0$ fixe mais inconnu.

Dans ce modèle, un estimateur classique $\tilde{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}^*$ est obtenu par la méthode dite des moindres carrés ordinaires (MCO) et est défini par

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Sauf mention contraire, nous supposerons par la suite que $p = p(n)$ est une fonction de n. Ce cadre théorique général permet notamment de décrire les situations pratiques où p n'est pas nécessairement négligeable devant n. On peut par exemple supposer que

$p(n) \to \infty$ plus ou moins vite lorsque $n \to \infty$ afin d'étudier le cas des données dites de grande dimension, pour lesquelles les approches classiques ne sont généralement pas recommandées et les approches pénalisées peuvent être préconisées. Si la matrice de design $\mathbf{X}$ est de rang p (ce qui implique notamment que $p \leq n$), on peut établir l'unicité de la solution $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Les propriétés théoriques de cet estimateur sont bien connues. En particulier, son *erreur de prédiction quadratique moyenne* est de l'ordre de

$$\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} = \mathcal{O}_{\mathbb{P}}\Big(\frac{p}{n}\Big).$$

Dans le cadre asymptotique « classique », où p est fixe et $n \to \infty$, ce résultat établit que l'erreur de prédiction quadratique moyenne tend vers 0 à la vitesse n^{-1} lorsque $n \to \infty$. Cependant, ce résultat établit également que l'estimateur des MCO souffre du *fléau de la dimension* : par exemple si $p = n^\alpha$, avec $0 < \alpha < 1$, l'erreur de prédiction moyenne ne tend plus vers 0 à la vitesse n^{-1}, mais à la vitesse $n^{-(1-\alpha)}$. Ce phénomène décrit parfois un *sur-ajustement aux données*. C'est notamment le cas lorsque $p = n$ et $\mathbf{X} = \mathbf{I}_n$, c'est-dire dans la version tronquée du modèle de suites gaussiennes : $Y_i = \beta_i^* + \varepsilon_i$, pour $i \in [n]$, avec $\beta_i^* \in \mathbb{R}$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $0 < \sigma^2 < 1$. L'estimateur des MCO y est donné par $\tilde{\boldsymbol{\beta}} = \mathbf{Y}$: les espérances β_i^* sont donc chacune estimées par chacune des observations Y_i et

$$\mathbb{E}\left\{\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n}\right\} = \mathbb{E}\left\{\frac{\|\mathbf{Y} - \boldsymbol{\beta}^*\|_2^2}{n}\right\} = \mathbb{E}\left\{\frac{\|\boldsymbol{\varepsilon}\|_2^2}{n}\right\} = \sigma^2.$$

Avec l'estimateur des MCO, l'espérance de l'erreur de prédiction quadratique moyenne ne tend pas vers 0 sous ce modèle.

Le fléau de la dimension n'est pas spécifique à l'estimateur des MCO. Il concerne la plupart des procédures d'estimation classique (estimation paramétrique et non paramétrique confondues), mais aussi les procédures de test, etc. Nous renvoyons le lecteur au chapitre introductif du livre de [Giraud, 2014] où le fléau de la dimension est illustré dans différentes situations.

A.2 La sélection de variables et les approches type BIC

Heureusement en pratique, la dimension *sous-jacente* est généralement bien plus faible que ne le laisse présager la dimension de la matrice de design. En effet, le vecteur de paramètres $\boldsymbol{\beta}^* \in \mathbb{R}^p$ a le plus souvent une certaine structure et peut être décrit par un nombre p_0 de paramètres souvent négligeable devant p : $p_0 \ll p$. Par exemple, les p variables X_j sont rarement toutes liées à la variable réponse Y. Ainsi, en notant $J^* = \{j \in [p] : \beta_j^* \neq 0\}$ et $p_0 = |J^*|$ le cardinal de J^*, on a typiquement $p_0 \ll p$ et le vecteur $\boldsymbol{\beta}^*$ est alors dit creux, parcimonieux ou sparse. Pour tout sous-ensemble $J \subseteq [p]$, et toute matrice $\mathbf{U}$ de dimension $n \times p$, notons $\mathbf{U}_J$ la matrice de dimension $n \times |J|$ constituée des colonnes de la matrice $\mathbf{U}$ d'index appartenant à J. Pour tout vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$, on note de même $\boldsymbol{\beta}_J$ le vecteur de $\mathbb{R}^{|J|}$ constitué des composantes de $\boldsymbol{\beta}$ d'index appartenant à J. Enfin, on note $J^c = [p] \setminus J$ le complémentaire de J dans $[p]$. Si J^* était connu, il suffirait d'appliquer les

MCO sur les données $(\mathbf{Y}, \mathbf{X}_{J^*})$ pour obtenir l'estimateur $\acute{\boldsymbol{\beta}}_{J^*}$ de $\boldsymbol{\beta}^*_{J^*}$. En posant $\acute{\boldsymbol{\beta}}_{J^{*c}} = \mathbf{0}_{p-p_0}$, on en déduirait l'estimateur $\acute{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}^*$ pour lequel l'erreur de prédiction quadratique moyenne serait sensiblement meilleure que celle correspondant à $\tilde{\boldsymbol{\beta}}$, à savoir $\mathcal{O}_{\mathbb{P}}(p_0/n)$ contre $\mathcal{O}_{\mathbb{P}}(p/n)$. Cependant, l'ensemble J^* des covariables pertinentes n'est en général pas connu et l'approche décrite ici n'est donc pas applicable en pratique. Elle suggère néanmoins de s'intéresser au problème classique de la sélection des variables pertinentes. Outre son intérêt évident quant à l'interprétation du modèle (lorsqu'on cherche à identifier les facteurs de risque d'une pathologie par exemple), la sélection des variables pertinentes peut conduire à des performances prédictives améliorées, et plus généralement des estimations plus précises, lorsque le vecteur $\boldsymbol{\beta}^*$ est effectivement creux, comme l'illustrent les résultats de la Figure A.1.

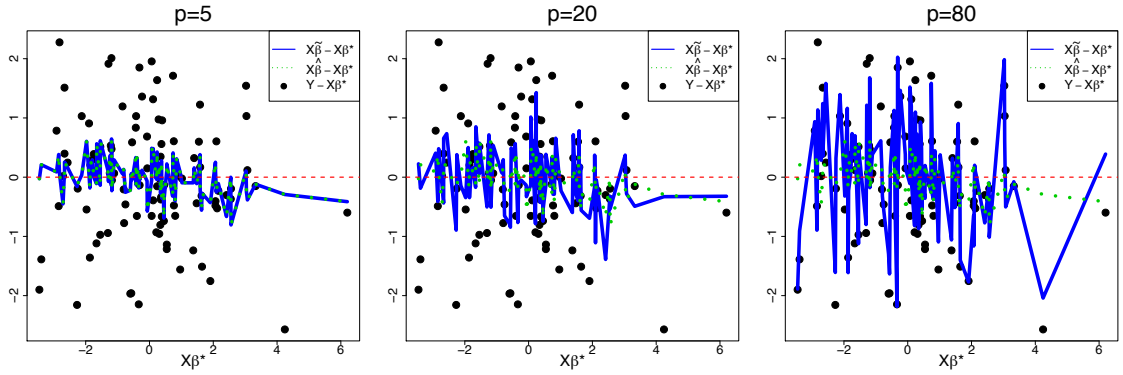


FIGURE A.1 – Illustration du fléau de la dimension dans le modèle de régression linéaire. On génère des données sous le modèle $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$, où $\mathbf{X}_0$ est une matrice de design de taille 100×5, $\beta_{0,j} = 0.8$ pour $j \in [5]$ et les ε_i sont $\mathcal{N}(0,1)$ pour tout $i \in [100]$. Pour $p = 5, 20$ et 80, on ajoute $p-5$ colonnes à la matrice $\mathbf{X}_0$ pour créer la matrice $\mathbf{X}$, de dimension $n \times p$, et on cherche à estimer $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_0^T, \mathbf{0}_{p-5}^T)^T \in \mathbb{R}^p$. L'estimateur $\tilde{\boldsymbol{\beta}}$ est celui des MCO, alors que $\hat{\boldsymbol{\beta}}$ inclut une étape de sélection de variables (il correspond à l'estimateur retourné par la procédure de sélection pas-à-pas ascendante utilisant le critère BIC). La variance des prédictions « centrées » augmente avec p pour l'estimateur des MCO (elle est comparable à celle des observations centrées pour $p = 80$ ici), alors qu'elle reste stable pour l'estimateur $\hat{\boldsymbol{\beta}}$.

La sélection des variables pertinentes est souvent une étape essentielle des analyses statistiques en épidémiologie et en recherche clinique. Dans ces disciplines, la question d'intérêt principal consiste généralement à établir le lien (ou l'absence de lien) entre la variable réponse Y et les covariables X_j, ou certaines d'entre elles après ajustement sur les autres. Pour ce faire, les procédures classiques comprennent les tests de comparaison, mais aussi certaines approches reposant sur des critères pénalisés, tel que le BIC [Schwarz et al., 1978]. Soit $J \subseteq [p]$, et $\tilde{\boldsymbol{\beta}}^{(J)}$ l'estimateur des MCO obtenus sur les données $(\mathbf{Y}, \mathbf{X}_J)$ (i.e. en se restreignant aux covariables contenues dans J). On peut définir un critère BIC correspondant

à ce modèle [Kim et al., 2012],

$$\mathrm{BIC}(J) = \frac{\|\mathbf{Y} - \mathbf{X}_J \tilde{\boldsymbol{\beta}}^{(J)}\|_2^2}{2} + |J|\sigma^2 \frac{\log n}{2}.$$

Une manière classique de sélectionner les variables, et d'estimer les effets associés, consiste à déterminer le sous-ensemble $\tilde{J}$ et l'estimateur des MCO $\tilde{\boldsymbol{\beta}}^{(\tilde{J})}$ correspondant tel que $\mathrm{BIC}(\tilde{J})$ soit minimal. En notant $\lambda = \sigma^2 \log(n)/2$ et en remarquant que $|J| = \|\tilde{\boldsymbol{\beta}}^{(J)}\|_0$ où $\|\cdot\|_0$ est la « norme » L_0, $\|\mathbf{x}\|_0 = |\{j \in [p] : x_j \neq 0\}|$, ce problème de sélection de variables revient finalement à résoudre le problème d'optimisation suivant :

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda \|\boldsymbol{\beta}\|_0 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p. \tag{A.2}$$

En notant $\tilde{\boldsymbol{\beta}}^{\mathrm{BIC}}$ une solution de ce problème d'optimisation, l'ensemble des variables sélectionnées par la procédure est alors $\tilde{J} = \{j \in [p] : \tilde{\beta}_j^{\mathrm{BIC}} \neq 0\}$ et leurs effets estimés sont contenus dans $\tilde{\boldsymbol{\beta}}_{\tilde{J}}^{\mathrm{BIC}}$.

Le problème d'optimisation (A.2) correspond à une version pénalisée, par la norme L_0 du vecteur de paramètres, de celui résolu dans les MCO. Ce critère est la somme de deux quantités : la première mesure l'adéquation aux données, alors que le second pénalise plus ou moins fortement les vecteurs $\boldsymbol{\beta} \in \mathbb{R}^p$: ces vecteurs sont d'autant plus pénalisés que leur support $J = \{j \in [p] : \beta_j \neq 0\}$ est de cardinal $|J| = \|\boldsymbol{\beta}\|_0$ élevé. En pénalisant les vecteurs à grand support, le critère BIC encourage les vecteurs creux, et permet ainsi d'opérer une sélection des variables. La consistance en sélection de variable est garantie sous certaines conditions ; voir par exemple [Kim et al., 2012].

Le critère BIC est très utilisé en pratique. Cependant le problème d'optimisation (A.2) est non convexe et ne peut donc pas être résolu « rapidement ». La résolution numérique de ce type de problème est dite combinatoire puisqu'il n'existe en général pas d'autres approches que celle consistant à énumérer l'ensemble des solutions possibles (ici les 2^p modèles qui correspondent à l'ensemble des parties de $[p]$), construire les modèles, calculer les critères BIC et renvoyer le modèle correspondant au critère BIC minimal. Dès lors que $p \geq 30$, il n'est pas raisonnable d'énumérer les 2^p modèles. Pour utiliser le BIC en de tels cas, on le combine le plus souvent à des heuristiques qui permettent de ne parcourir qu'un sous-ensemble des 2^p modèles. Les plus utilisées, en épidémiologie et recherche clinique en tout cas, sont les approches dites pas-à-pas (*stepwise* en anglais), qui peuvent être ascendantes ou descendantes, voire hybrides [Hocking, 1976].

A.3 Relaxation convexe du critère BIC : le lasso

Pour résumer, en pénalisant le critère des MCO par la norme L_0 du vecteur des paramètres du modèle, on obtient un critère de type BIC qui opère une sélection des variables. Cette sélection est consistante, sous certaines hypothèses, mais le minimum global du critère est difficile à obtenir numériquement, sauf à considérer des cas où p est très petit. Depuis une vingtaine d'années, la recherche en statistique s'efforce à proposer des critères pénalisés

alternatifs, qui soient simples à résoudre numériquement tout en renvoyant des estimateurs présentant de bonnes propriétés statistiques [Candes and Tao, 2007, Tibshirani, 1996, Fan and Li, 2001, Bühlmann and van de Geer, 2011, Giraud, 2014]. D'une manière générale, on peut en effet voir le critère en (A.2) comme un cas particulier du critère suivant :

$$\frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + P_\lambda(\boldsymbol{\beta}) \tag{A.3}$$

où $P_\lambda : \mathbb{R}^p \to \mathbb{R}$ est une fonction dépendant d'un paramètre $\lambda \geq 0$, dite de pénalité. Le critère en (A.2) est obtenu avec le choix $P_\lambda = \lambda\|\boldsymbol{\beta}\|_0$, mais de nombreux autres choix ont été proposés et étudiés dans la littérature, pour encourager certains vecteurs en fonction des caractéristiques des données traitées. Un choix populaire qui a attiré une attention particulière tant dans la littérature théorique qu'appliquée, est le lasso décrit dans [Tibshirani, 1996]. Il consiste à remplacer la norme L_0 du BIC par son enveloppe convexe sur l'intervalle $[-1, 1]$ [Jojic et al., 2011]. Celle-ci correspond à la norme L_1, et le lasso utilise donc la pénalité $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, où $\|\boldsymbol{\beta}\|_1 = \sum_{j\in[p]} |\beta_j|$ est la norme L_1 du vecteur $\boldsymbol{\beta}$. En utilisant cette relaxation, le problème d'optimisation qui en résulte, à savoir

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{A.4}$$

est convexe. Il se résout numériquement par des méthodes d'optimisation convexe, dont les complexités algorithmiques sont typiquement polynomiales en p (et non plus exponentielles) et en n [Boyd and Vandenberghe, 2004]. De nombreux algorithmes sont disponibles dans des packages du logiciel R notamment (`lars`, `glmnet`, `penalized`, etc.). Certains d'entre eux (`lars` notamment ; voir [Efron et al., 2004]) sont particulièrement adaptés pour déterminer l'ensemble des solutions $\hat{\boldsymbol{\beta}}(\lambda)$ pour toutes les valeurs possibles du paramètre λ, ce qu'on appelle le *regularization path*. Ceci revêt un intérêt particulier en pratique puisque ce paramètre λ doit être choisi avec précaution, généralement en fonction des données ; voir le paragraphe A.5 ci-dessous. En particulier, pour tout $\lambda > 0$, les solutions du problème (2.1) sont typiquement creuses et en notant $\hat{J}(\lambda) = \{j \in [p] : \hat{\beta}_j(\lambda) \neq 0\}$, il a été établi que $\hat{J}(\lambda) = J^*$ avec grande probabilité pour un choix approprié du paramètre de pénalité λ, et ce sous des hypothèses portant sur la matrice de design $\mathbf{X}$, le support J^* de $\boldsymbol{\beta}^*$ et la « force du signal » (mesurée par $\beta^*_{\min} = \min_{j\in J^*} |\beta^*_j|$) [Zhao and Yu, 2006, Zou, 2006, Wainwright, 2009]. Le lasso est alors dit consistant en sélection de variables, ou *sparsistent*. L'hypothèse principale portant sur la matrice de design est celle dite d'irrepresentabilité (*irrepresentability condition*), qui stipule que $\Lambda_{\min}(\mathbf{X}_{J^*}^T\mathbf{X}_{J^*}) > 0$ et

$$\max_{j\notin J^*} \|(\mathbf{X}_{J^*}^T\mathbf{X}_{J^*})^{-1}\mathbf{X}_{J^*}^T X_j\|_1 < 1. \tag{A.5}$$

Autrement dit, la condition d'irreprésentabilité assure que le modèle restreint à J^* est identifiable et que les colonnes de J^{*c} ne sont pas trop alignées sur celles de J^*. Sous des hypothèses moins restrictives sur la matrice de design $\mathbf{X}$, on peut montrer [Bickel et al., 2009, Dalalyan et al., 2014] que l'erreur de prédiction quadratique moyenne est *oraculaire*, de l'ordre de $\mathcal{O}_{\mathbb{P}}(p_0 \log(p)/n)$ lorsque $\|\boldsymbol{\beta}^*\|_0 = p_0$: au terme $\log(p)$ (ainsi qu'aux constantes)

près, c'est la vitesse que l'on obtiendrait pour l'estimateur des MCO restreint aux variables de J^*.

Ainsi, le lasso combine de bonnes propriétés numériques et, sous certaines hypothèses, de bonnes propriétés statistiques (consistance en sélection de variables, erreur de prédiction oraculaire). Il n'est cependant et bien sûr pas parfait puisque les hypothèses assurant ses bonnes propriétés sont à la fois fortes et difficiles voire impossibles à vérifier sur les données. D'autre part, le lasso renvoie des estimations typiquement biaisées. C'est l'effet de *shrinkage* : pour $\lambda > 0$, chaque composante non nulle du vecteur solution du problème (A.4) fournit généralement une estimation dont la valeur absolue est ramenée vers 0 par rapport à la composante correspondante de $\boldsymbol{\beta}^*$. Diverses extensions du lasso ont été proposées, notamment pour réduire ces biais.

A.4 Extensions du lasso

La version OLS-Hybrid du lasso [Efron et al., 2004] consiste, pour toute valeur λ, à ré-estimer les composantes non-nulles du vecteur $\hat{\boldsymbol{\beta}}(\lambda)$ solution du lasso. Pour ce faire, et si $\hat{J}(\lambda) = \{j \in [p] : \hat{\beta}_j(\lambda) \neq 0\}$ n'est pas trop grand, on utilise la méthode des MCO en se restreignant aux variables contenues dans $\hat{J}(\lambda)$. La ré-estimation étant faite sans pénalité, les biais du lasso sont éliminés, mais d'autres types de biais peuvent apparaître du fait de la sélection de variables préalable à l'étape d'estimation par MCO [van Houwelingen and Le Cessie, 1990].

Une généralisation du lasso OLS-Hybrid est le lasso relaxé [Meinshausen, 2007]. Celui-ci dépend d'un deuxième paramètre $0 \leq \phi \leq 1$, qu'on appellera ici paramètre de relaxation. Etant donnée une solution $\hat{\boldsymbol{\beta}}(\lambda)$ du lasso obtenue pour le paramètre de pénalité λ, le lasso ϕ-relaxé consiste à résoudre le problème d'optimisation suivant

$$\text{minimiser} \quad \frac{\|\mathbf{Y} - \mathbf{X}_{\hat{J}(\lambda)}\boldsymbol{\beta}\|_2^2}{2} + \phi\lambda\|\boldsymbol{\beta}\|_1 \quad \text{sur } \boldsymbol{\beta} \in \mathbb{R}^{|\hat{J}(\lambda)|}, \tag{A.6}$$

où $\hat{J}(\lambda)$ est le support de $\hat{\boldsymbol{\beta}}(\lambda)$. En d'autre terme, le lasso ϕ-relaxé consiste à résoudre le lasso avec le paramètre de pénalité diminué, égal à $\phi\lambda \leq \lambda$, en se restreignant aux covariables du support $\hat{J}(\lambda)$ de $\hat{\boldsymbol{\beta}}(\lambda)$. Les biais du lasso dépendant du paramètre de pénalité, le lasso ϕ-relaxé a pour vocation de réduire ces biais. Le lasso 0-relaxé (avec $\phi = 0$) revient à la version OLS-Hybrid du lasso. Le lasso 1-relaxé revient quant à lui au lasso. Dans un cadre asymptotique en n (avec $p = p(n)$), Meinshausen établit notamment que l'erreur de prédiction du lasso relaxé converge plus rapidement vers 0 que celle du lasso, sous certaines hypothèses.

Parmi les autres approches corrigeant le biais des estimateurs lasso, on peut également citer le lasso adaptatif de [Zou, 2006], ou encore le lasso itéré de [Candes et al., 2008]. Tous deux remplacent la norme L_1 par une version pondérée de celle-ci : $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j\in[p]} w_j|\beta_j|$. Les poids w_j dépendent directement d'estimations initiales des composantes β_j^*. Le principe général est de pénaliser plus fortement les composantes dont les estimations initiales sont faibles en valeur absolue, et moins fortement les composantes correspondant à des estimations élevées (réduisant ainsi les biais sur ces variables). Dans le cas où p est fixe et

$n \to \infty$, [Zou, 2006] établit notamment la consistance en sélection de variables lorsque les poids sont de la forme $w_j = |\breve{\beta}_j^{-1}|$ si $\breve{\beta}_j$ est un estimateur $\sqrt{n}$-consistant de β_j^*, et ce sous des hypothèses très générales sur la matrice de design (en particulier, sans faire d'hypothèse d'irreprésentabilité). [Candes et al., 2008] proposent quant à eux d'utiliser des poids de la forme $w_j = 1/|\hat{\beta}_j(\lambda^{\mathrm{CV}}) + \epsilon|$, avec ϵ petit et λ^{CV} le paramètre de régularisation sélectionné par cross-validation après un premier lasso standard.

A.5 Calibration du paramètre de régularisation

Les résultats théoriques pour le lasso, et les approches pénalisées en général, dépendent en particulier du choix du ou des paramètre(s) de régularisation. Leur valeur optimale dépend elle-même généralement de quantités inconnues : la variance du bruit dans le cas gaussien ainsi que certaines « constantes » qui dépendent de la matrice de design et de la structure, inconnue, du vecteur $\boldsymbol{\beta}^*$ par exemple. En pratique, une étape essentielle lors de l'application de ces approches est donc la sélection, ou calibration, des paramètres de régularisation optimaux. Deux familles de critères sont le plus souvent utilisées pour opérer cette sélection. Premièrement, on peut utiliser des méthodes de ré-échantillonnage (validation croisée, etc.) pour estimer l'erreur de prédiction moyenne associée à chaque choix particulier des paramètres de régularisation, et sélectionner ceux qui minimisent ce critère. Deuxièmement, on peut utiliser les critères tels que le BIC. Quelque soit le critère retenu, on peut le calculer soit directement à partir des estimations retournées par l'approche pénalisée considérée, soit en ré-estimant les paramètres sans pénalité, mais sous la contrainte induite par la structure du vecteur retourné par l'approche pénalisée (via la version OLS-Hybrid du lasso par exemple, ou une extension idoine).

Le choix de la méthode de sélection des paramètres de régularisation dépend à la fois de la finalité de l'analyse statistique (construction d'un modèle prédictif ou sélection des variables pertinentes), et du ratio n/p. En particulier, si la sélection des paramètres pertinents est la question d'intérêt principal, la validation croisée, sans ré-estimation, ne permet généralement pas de sélectionner le bon modèle [Meinshausen and Bühlmann, 2006, Meinshausen and Bühlmann, 2010, Wang et al., 2007]. Les critères obtenus après ré-estimation, et en particulier les critères de types BIC, sont mieux adaptés à cette situation, notamment si le ratio n/p est assez grand [Meinshausen, 2007].

Dans la plupart des applications auxquelles j'ai été confronté en épidémiologie, la question d'intérêt principal est celle de la sélection des variables pertinentes. D'autre part, elles se plaçaient le plus souvent dans un cadre où le ratio n/p n'était pas petit. Ainsi, et sauf mention contraire, la sélection des paramètres de régularisation est effectuée dans ce manuscrit par minimisation d'un critère de type BIC, après ré-estimation des paramètres. Ce type de critère est désigné sous le terme générique 2stepBIC dans ce manuscrit.